Automation in Construction 146 (2023) 104695

Contents lists available at ScienceDirect

Automation in Construction

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

Digital twin enabled fault detection and diagnosis process for building HVAC systems

Xiang Xie [a,b,c,*], Jorge Merino [b,c], Nicola Moretti [b,c], Pieter Pauwels [d], Janet Yoon Chang [b], Ajith Parlikad [b,c]

[a] School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
[b] Institute for Manufacturing, Department of Engineering, University of Cambridge, Cambridge, CB3 0FS, UK
[c] Centre for Digital Built Britain, University of Cambridge, Cambridge, CB3 0FA, UK
[d] Department of the Built Environment, Eindhoven University of Technology, Eindhoven, 5600 MB, Netherlands

ARTICLE INFO

Keywords:
Building intelligence
Digital twin
Fault detection and diagnosis
Semantic web
Data integration
Real-time data
Metadata tagging
Asset management

ABSTRACT

The emerging concept of digital twins outlines the pathway towards intelligent buildings. Although abundant building data carries an overwhelming amount of information, if not well exploited, the redundant and irrelevant data dimensions result in the overfitting problem and heavy computational load. Taking the fault detection and diagnosis process for building HVAC systems as the case, this paper adopts a symbolic artificial intelligence technique to identify informative sensory dimensions for building-specific faults by exploring the symbolic representation of labelled time-series. To preserve this ad-hoc temporal knowledge in the digital twin ecosystem, machine-readable fault tags are defined to label corresponding sensor entities. A digital twin data platform is developed to annotate the real-time data with fault tags and produce filtered low-latency data streams associated with a specified tag to automate this process. This paper describes a digital twin-based approach to automatically identify and pick up informative data to support dynamic asset management.

1. Introduction

People living in modern society spend a considerable amount of time indoors. For example, the US Environmental Protection Agency reported that Americans spend, on average, 87% of their time inside buildings [1]. Comparably, UK adults spend an average of 22 h a day, amounting to over 90% of their time, in enclosed spaces [2]. The orchestration of diverse building systems is imperative to create an optimal indoor environment in modern facilities. To support this, the recent digital transformation of buildings has pushed building operations from conventional and programmatic to responsive and intelligent [3]. Different building data from disparate sources allow complex building systems to be intelligent, leveraging ubiquitous sensing capability and computing power [4].

As one of the enablers of building intelligence, a digital twin is suitable for replicating, predicting, and optimising the conditions and behaviours of assets (i.e., buildings, systems and components) during their lifecycle. According to [5], the digital twin is a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometric level. In the Architecture, Engineering, Construction, and Operation (AECO) industry, Boje et al. [6] defined the ability of construction digital twin over the entire building lifecycle under the 'Virtual-Data-Physical' paradigm. To bring the digital twin concept into practice, Seghezzi et al. [7] defined an occupancy-based digital twin to monitor actual occupancy levels in the building to optimise space utilisation, customise cleaning activities and contracts. Lu et al. [8] implemented a digital twin-enabled asset monitoring solution to identify anomalies of critical assets. Jafari et al. [9] presented a digital architecture that assimilates data streams on asset conditions and behavioural patterns to drive the energy simulation through an analytic engine for the asset systems. O'Dwyer et al. [10] proposed a digital twin tool for coordinating multi-vector energy systems, considering subsystems like building heating systems independently, yet ensuring various localised systems adhere to high-level system constraints.

Until now, it remains an open-ended question how to implement digital twins for the distributed and dynamic building systems to gain efficiency and effectiveness during operation. Although digitisation can benefit building performance and overall business profitability, adopting digital technologies to help the AECO industry remains challenging due to the data-intensive asset management processes [11]. Fragmented information sources produce a significant amount of multi-modal data, static [e.g., Construction-Operations Building Information

* Corresponding author at: School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.
E-mail address: xiang.xie@ncl.ac.uk (X. Xie).

https://doi.org/10.1016/j.autcon.2022.104695
Received 3 June 2022; Received in revised form 27 October 2022; Accepted 28 November 2022
Available online 9 December 2022

Exchange (COBie) spreadsheets] and dynamic (e.g., Internet of Things devices, building management systems), that can be used to enable building intelligence. It is a common practice to decompose the building systems into 'subsystems' (i.e., 'divide-and-conquer' strategy) and gain localised insights from disaggregated datasets. It ensures efficient decision-making for asset management in a coordinated way under the digital twin analytical framework [12]. The disaggregation, as the ad-hoc knowledge aiming at picking up informative and representative data to drive the decision-making, needs to be incorporated into the digital twin ecosystem in a machine-readable way.

This study aims to demonstrate a digital twin enabled Fault Detection and Diagnosis (FDD) process, which aims at identifying abnormalities in building Heating, Ventilation and Air Conditioning (HVAC) systems to prevent poor indoor air quality, thermal discomfort, and low productivity [13]. With semantic web technologies [14], the integration of heterogeneous data from different sources and characteristics can be easily achieved. More importantly, to support more controlled decision-making for fault detection, 'knowledge tags' can be defined semantically to form fault-targeted subsystems, which encapsulate the most representative subset of data dimensions to the corresponding faults respectively. This is a trial to better utilise massive real-time data. Although data carries a bewildering amount of information, excess HVAC sensory data actually degrades FDD performance, masking and flushing the more informative dimensions and causing overfitting between targeted faults and irrelevant dimensions [15]. To reduce the risk of overfitting and avoid the 'curse of dimensionality' [16], a Bag-of-Words (BoW) based feature extraction and selection method is introduced to identify the most informative sensory data dimensions from the labelled data. The knowledge tags are defined to incorporate this contextual knowledge learned from real-time data and automatically update the semantic model. It is key to enabling the needed self-evolving character of the digital twin model and adaptively feeding appropriate data to realise FDD of the random faults concisely [17]. The proposed fault detection and diagnosis of building HVAC systems inform the development of digital twins and the awakening of building intelligence for effective asset management.

The rest of this paper is organised as follows. Section 2 includes the literature review of semantic web technologies used in building systems and the fault detection and diagnosis of building HVAC systems. Section 3 discusses the proposed methodology. Section 4 presents the case study, demonstrating the novel digital twin-enabled fault detection and diagnosis process for building HVAC systems. Finally, Section 5 presents a discussion followed by Section 6 concluding this study.

2. Literature review

2.1. Use of digital twin and semantics in asset management processes

By definition, the digital twin is a digital representation of physical assets, processes or systems in the built environment [18]. The built assets, inextricably linked into a complex and highly interconnected 'system of systems', deliver continuous services to various stakeholders [19]. The development of digital twins for built assets, where physical meets digital and sets out its core value proposition, aims to inform the decision makers of the system and improve their decisions driven by data, and ultimately deliver better outcomes for people.

Semantic web technologies play a key role in expressing, integrating and managing data and information for the development of building digital twin. As summarised in [20], the usage of semantics conventionally focused on: overcoming the interoperability issues amongst software tools and improving information exchange processes; connecting data across different domains, including but not limited to Building Information Model (BIM), Geographic Information System (GIS) and sensor data; and enabling logic-based declarative inference that extracts extra information from the original representations. These semantics-based models are the natural successor to the current ambition in BIM

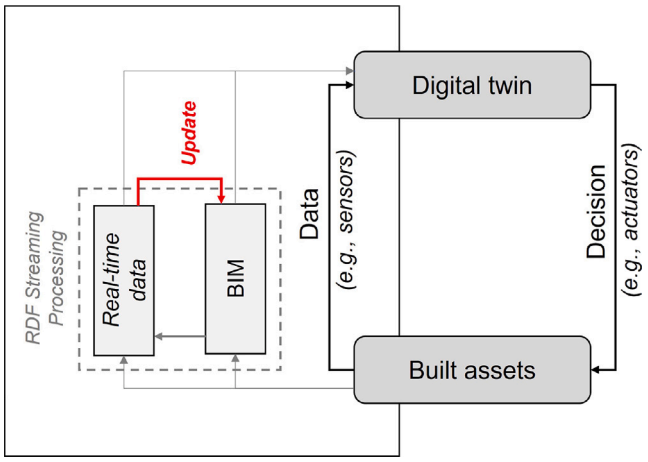

Fig. 1. Conceptual diagram of the digital twin analytical framework.

that has become pervasive in this industry since the 1980s [21]. In short, some of the values enriched by the semantic web based approach are: (1) the addition of a logical basis that allows declarative inference as opposed to procedural coding, and (2) the much more open and web-based approach to data integration.

What we highlight in this study is that, armed with semantics that is readily available in BIM models and more easily expressive in a graph-based semantic web world, digital twins can be created more flexibly with novel connections established using data associations continuously learned from real-time data [6]. Real-time data streams, for example those from sensors, carry a massive amount of information. To extract insights from the heterogeneous data streams, it is necessary to integrate them with solid knowledge in the form of taxonomies and class hierarchies defined by diverse domain ontologies. In this context, RDF Stream Processing (RSP) provides reasoning capabilities to infer implicit facts about Resource Description Framework (RDF) streams by incessantly answering SPARQL Protocol and RDF Query Language (SPARQL) queries (e.g., C-SPARQL, CQELS) [22,23]. Supporting algebraic operators such as queries, joints, filters, and aggregations, the RSP is demonstrated to be a useful engine in mediating the data pipelines from data sources to data storage, ultimately to applications and services on demand. However, the digital twinning process is an evolving process, as explained in Fig. 1. Deeper insights into the physical 'system of systems' are gradually and constantly upgraded through continuously ingesting the latest data streams. The up-to-date contextual knowledge acquired by various artificial intelligence algorithms [24] has got to be incorporated by introducing machine-readable connections and feeds back to the graph-based semantic models [25]. This is critical in the development of the digital twin analytical framework, because knowledge, initially held in domain ontologies and more importantly learned from latest data, must be expressed and used automatically to minimise the human intervention.

The focus of this paper is to experiment with the integration of ad-hoc knowledge learned through symbolic artificial intelligence techniques into the digital twin ecosystem. Aiming at reducing the data numerosity while preserving the most relevant information, this study enables decomposing complex systems into separated subsystems by defining knowledge tags in the corresponding semantic models. And this will be explained in the next subsection. In this way, the data analysis becomes more computationally efficient leveraging the divide-and-conquer strategy, which instinctively supports more localised decision-making.

2.2. Domain ontologies for buildings and knowledge tagging

To enable the integration of ad-hoc knowledge for supporting asset management processes, a stable semantic basis is needed to enable

a solid backbone for the digital twins. In this regard, a series of domain-specific ontologies have been developed to solve problems in different application domains. Reviews of diverse domain ontologies are presented in Zhong et al. and Pauwels et al. [20,26,27] and out of the scope for this paper. Typically, two large clusters of research can be found, one situated in the conventional AEC domain, and the other in the building HVAC systems domain.

In the AEC domain, most of the research on semantic modelling of buildings is situated around the Industry Foundation Classes (IFC) and Linked Building Data (LBD). The Web Ontology Language (OWL) implementation of IFC is summarised in [28,29]. And under the umbrella of LBD, several lightweight domain ontologies have been created using OWL, such as the Building Topology Ontology (BOT) [30,31]. Starting from the BOT ontology, several extended domain ontologies are developed, aiming to describe building products/elements (e.g., walls, windows, devices) and properties [32–35]. In the building HVAC systems domain, much less focus is put on the actual building topology, nor on the specific building products or their properties. Instead, focus is concentrated on the representation of the systems and sensing points as well as generated time-series data. The Brick ontology[1] and Haystack tagging ontology[2] have been developed to collect specific object types and properties for building system components [36–38]. Of particular difficulty in this domain are flows, devices states, control logics, and sensor data streams. Kukkonen et al. [39] designed a Flow Systems Ontology[3] (FSO) for describing the composition of flow systems, such as the chilled beam system, and their mass and energy flow relationships. Xie et al. [40] proposed a federated ontology for representing building spatial and metering system hierarchies, which connects building submetering data with spatial characteristics for fine-grained energy analysis.

By definition, a tag is a keyword or term deliberately assigned to a piece of information. Based on the structured semantics in the ontologies, the semantic tagging binds sets of 'knowledge tags' to entities or classes, helping to reserve, browse and search tacit knowledge of domain experts or particularly ad-hoc temporal knowledge learned from data [41]. The captured knowledge shows in the forms of descriptions, categorisations, classifications, comments, notes, hyperdata, hyperlinks, or references that are collected in tag profiles. Mishra et al. [25] proposed to automate Haystack tagging by leveraging rule-based knowledge (e.g., inferred semantic facts based on raw point names) and data-driven techniques (e.g., supervised labelling using time-series data). In this study, the Brick ontology with relatively basic and lightweight conceptual structures is adopted, allowing for the flexible description of points within a building and the formalisation of mapping between tags and entities/classes. The extra information resulting from introduced tags, adds additional value, context, and meaning to the explicitly represented information.

2.3. Fault detection and diagnosis of building HVAC systems

Fault detection and diagnosis, like many other intelligent functions, allows to make better-informed decisions based on the multifaceted insights gained from massive data from building automation systems (BAS). The insights can either be acquired from calibrated physical models or be mined using data-driven approaches [13]. Accurate modelling of the HVAC system is vital for the physical model-based FDD. However, considering the interacting subsystems with complicated and non-linear dynamics, it is challenging and expensive to approximate the system's characteristics using the first principle model [42] or grey box model [43]. Furthermore, the uncertainties from inaccurate model formulations or modifications to the original system would greatly compromise the performance of model-based FDD.

Alternatively, data-driven approaches are widely explored to realise the FDD of HVAC systems. Relying on historical and online data, data-driven FDD reveals the intra-attribute patterns, which help identify the normal and faulty system behaviours in a supervised or unsupervised way. For instance, Zhao et al. [44,45] developed a method based on diagnostic Bayesian networks (DBNs) for diagnosing 28 types of faults in air handling units (AHUs) in buildings; Yu et al. [46] and Petrova et al. [47] adopted association rule mining (ARM) to discover faults in air conditioning systems from building operational data, and in the second case returning the found association rules back into the core semantic model.

Data-driven FDD is relatively simple to implement in an automated manner. However, excessive trials on adopting data-driven FDD in actual buildings hardly give satisfactory results. The unsatisfactory performance usually results from the fact that the performance of FDD largely depends on the quality of the data input. The selection of the most relevant data dimensions from the complete dataset, as the key knowledge to support data-driven FDD, is vital to decrease the risks of overfitting and reduce computational complexity. Yan et al. [48] and Mulumba et al. [49] used a filter-based algorithm called ReliefF to select the optimal feature subset for the FDD application to chillers. Li et al. [50] adopted the information greedy feature filter (IGFF) to eliminate noisy and noninformative features that compromise the FDD, maximising mutual information between selected features and the fault labels. Zhang et al. [51] proposed a statistical feature extraction technique (standard deviation, mean, minimum, maximum) with varying window sizes and use the filter and wrapper methods to select the optimal dataset based on cross-validation. To sum up the above literature, some focused on the typical faults of specific equipment (e.g., condenser fouling of chiller), preserving the physical significance of the selected features [52] but not necessarily guaranteeing equivalent performance when utilised in alternative scenarios. Similar to the idea of IGFF and statistical measures, in this study, a lightweight Bag of Words (BoW) based feature extraction and selection method [53] is adopted to recognise the most distinguishing sensory dimensions based on the time-series features. The reality is, in addition to typical faults, buildings are prone to many 'baffling' faults. The BoW-based method is appropriate in determining a selection of the nominated features for intractable faults in complicated systems. Besides, compared with IGFF and statistic measures, the BoW-based method reduces the size of a time series without losing key information, while also enjoying higher tolerance of sensor noise. The identified sensory dimensions, forming separated subsystems for detecting corresponding faults, are preserved through fault tagging and annotation. This piece of new knowledge helps to deliver FDD functionality using the most relevant dimensions and adapt the data pipelines accordingly.

3. Implementation

3.1. Digital twin enabled FDD process

Different from the general data-driven fault detection and diagnosis studies, this paper focuses on the realisation of the FDD process in a systematic way. By recognising the subsystems informative to arbitrary faults and labelling the relevant sensors within each subsystem using the knowledge tags, the ad-hoc temporal knowledge is preserved to customise corresponding data pipelines to drive FDD. It is key to automating the digital twin enabled HVAC FDD process, illustrated in Fig. 2. More specifically, facility management professionals are encouraged to name the most concerned or frequent faults in a particular building HVAC system, and the corresponding work orders can be retrieved from computer-aided facility management (CAFM) software, containing the historical duration of these faults. A case-specific fault watch list is established to record all concerned faults raised by facility managers, and normal and faulty data can be acquired according to the work orders for faults in the watch list. Importantly, for any particular

[1] https://brickschema.org/schema/Brick
[2] https://project-haystack.org/doc/docHaystack/Ontology
[3] https://alikucukavci.github.io/FSO/

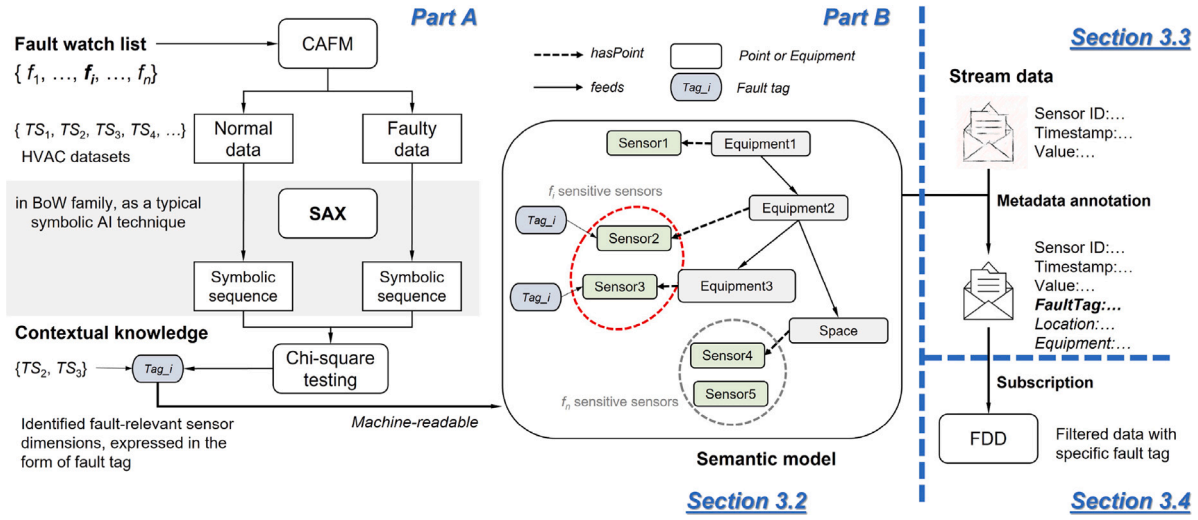

Fig. 2. Diagram of fault detection and diagnosis process for building HVAC systems.

fault, it is unnecessary to utilise the whole HVAC dataset for enabling the FDD. With the help of the supervised BoW-based feature extraction and selection method, sensor collections restricted to detect every nominated fault can be uniquely selected based on the labelled normal and faulty data, automatically forming the subsystem dedicated to the particular fault without the need for human intervention. Fault tags are created in the semantic model accordingly, integrating these pieces of knowledge into the digital twin ecosystem as ad-hoc knowledge to associate the most relevant sensors with the corresponding fault tags. The continuously produced data from sensors is annotated with the associated fault tags, making sure that the processed data is appended to be independent and self-contained. The data streams in the form of a sequence of JavaScript Object Notation (JSON) objects are filtered to find out those containing a particular fault tag, and the occurrence of the fault can be alarmed based on the Goodness-of-Fit test.

In summary, the digital twin analytical framework automatically identifies the most valuable real-time data dimensions for specific faults, according to the knowledge learned from data through the BoW-based feature selection method (Section 3.2-part A). By semantically defining fault tags to label these picked dimensions (Section 3.2-part B), this knowledge is preserved to allow the automation of the FDD functionality, processing the continuous real-time data streams without human intervention and fetching the right data for the nominated fault types in the customised fault watch list (Section 3.3). The subset of real-time data streams, corresponding to those fault-relevant dimensions, is used to detect the occurrence of specific faults. The fault alarm is triggered once the frequency of the symbols transformed from the latest data deviates from the normal condition based on the criterion of Kullback–Leibler divergence (Section 3.4). Adopting the 'divide-and-conquer' strategy, this framework promotes asset management efficiency by making more controlled decisions using only the most informative and representative subset of data.

3.2. Brick schema and fault tagging for FDD

The Brick schema (version 1.2) is adopted as the ontology for establishing the semantic model for building systems, because of its comprehensive expressiveness and wide adoption in describing the HVAC system components and their connection with monitored sensory data. It provides standardised vocabularies for representing the physical, logical, and virtual assets in buildings and the relationships between them [36]. Specifically, `point`, `location`, and `equipment` are defined as the scaffolding of Brick's class hierarchy [27]. It has been verified to be capable of semantically twinning HVAC systems in buildings, with standardised metadata of HVAC entities and their relationships [37].

To further extend the expressiveness of the Brick model in contextual knowledge, knowledge tags for faults are defined as atomic facts under the Brick Tag namespace to reserve the most relevant sensory dimensions effective in detecting specific faults. Standard semantic definitions for the HVAC system faults can be integrated to orderly classify the possible HVAC faults according to their characteristics and causal relations [54]. The incorporated fault taxonomy can not only unify the naming convention of various faults, but more importantly, define a consistent physical hierarchy, which can be used to classify faults occurring at different levels of operation (e.g., component level, sub-system level, whole system level). However, in practice, there exist faults with undetermined causes. In this paper, the fault taxonomy is not used for better flexibility in accommodating any customised fault type. Each proposed fault tag is associated with a set of selected sensors (i.e., points) accordingly. And the connections in the form of *hasTag/isTagOf* (defined in Brick schema) are introduced to map between each type of fault and the associated sensors. This can reduce the computational complexity of the FDD functionality and facilitates the identification of each fault independently. In the case of the FDD for building HVAC systems, the Brick ontology is adopted instinctively. But the knowledge tagging should apply to any other suitable ontologies, as a 'carrier' of knowledge.

The relevant sensory dimensions are selected through analysing the symbolic aggregate approximation (SAX) of labelled data. This assumes that a considerable number of sensors is embedded in the HVAC systems, generating a massive amount of data. The FDD process is challenging because the intricate dependencies between fault-irrelevant sensory data dimensions unnecessarily increase the problem complexity and mask the true associations between the multisensor data and the targeted HVAC faults. In this study, the Bag-of-Word (BoW) based feature extraction and selection method is adopted to identify the fault-relevant dimensions using labelled normal and faulty data, and thus support subsystem disaggregation for detecting targeted HVAC faults [17]. The SAX in the BoW family is used, for extracting features (codewords) from the HVAC sensory data. Because sensor data in buildings is typically sampled with the interval of 1 min, 5 min, 15 min or even an hour, it is believed that building systems always stay in a pseudo-steady state. As a generalised version of standard deviation, mean and other statistical measures, the temporal characteristics extracted using SAX implicitly carry important information indicating, for example, the occurrence of specific events.

Leveraging symbolic artificial intelligence, the SAX has discretisation functions that transform data segments into symbols and further

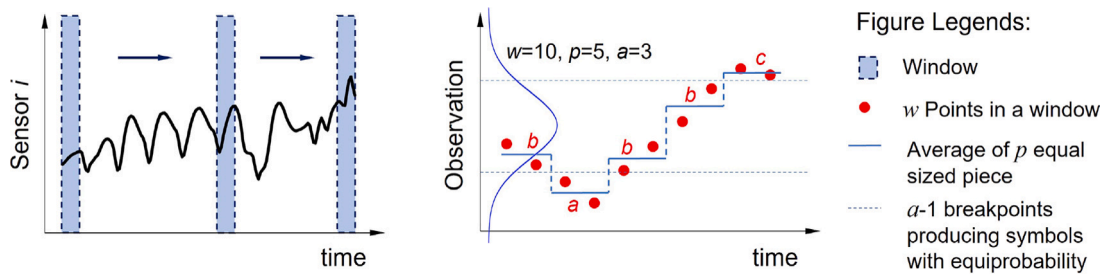

Fig. 3. Illustration of converting time-series to codewords with SAX.

transform each time-series to codewords (i.e., symbolic aggregates). A codeword is a combination of unordered alphabetic letters, like '*babbc*'. As shown in Fig. 3, for the normalised time-series data from the sensor i ($\boldsymbol{x}_i = [x_{1i}, x_{2i}, x_{ti}]^T$), a sliding window with length w ($w \ll t$) is used to segment the ith sensory data into fixed-length observations in an overlapping manner. Assuming the kth observation as $[x_{(t_k-w+1)i}, x_{(t_k-w+2)i}, \dots, x_{t_k i}]^T$, the observation is further partitioned into p equal sized pieces, with p denoting the codeword length. These p pieces are discretised into p letters, from one of the a alphabets. To achieve maximum entropy partitioning, the breakpoints of the discretisation (corresponds to the horizontal dashed lines in Fig. 3) are defined to produce a alphabets with equiprobability under a Gaussian curve [55]. If the average value of w/p observations is below the smallest breakpoint, these pieces are mapped to the symbol '*a*'. If greater than or equal to the smallest breakpoint and less than the second smallest breakpoint, they are mapped to the symbol '*b*', and the procedure continues up to the highest average value that is mapped to the last symbol of the chosen alphabet.

The frequency of the symbols in codewords can evidence repetitive patterns in the data. Therefore, the histograms of codewords extracted for normal and faulty data respectively, are used as the criteria for feature selection. The χ^2 test is a typical statistical hypothesis test, determining whether there exists a statistically significant difference between two categorical variables. Here, the χ^2 test is used to test the consistency/distinguishability of codeword histograms extracted for normal and faulty operational data, and identify codewords that are most relevant to faults. Let T and P denote the counts of a specific codeword and all codewords found in faulty data, and M and Z denote the counts of a specific codeword and all codewords found in normal and faulty data together. The χ^2 score of the corresponding codeword is calculated as [17]:

$$\chi^2 = \frac{Z(TZ - MP)^2}{PM(Z-P)(Z-M)} \tag{1}$$

The χ^2 score follows a standard chi-square distribution ($\chi^2 \sim \chi_1^2(0)$). Only those codewords with χ^2 score passing a predefined threshold are kept as the distinguishing features that can differentiate normal and faulty HVAC system conditions. Those sensory data dimensions that generate these distinguishing codewords (normal versus faulty) are obviously more relevant to the specific HVAC fault. Accordingly, the defined fault tag is then connected with these sensors through *hasTag/isTagOf*, forming unique subsystems for detecting customised faults.

3.3. Incorporation of fault tags in the digital twin data platform

This section addresses the integration of knowledge tags defined in the Brick model and real-time data typically from building management systems (BMS) or IoT sensors. Metadata, including the knowledge tags here, means 'data about data' and can be defined as pieces of information describing entities in buildings. Real-time data, referred to as stream data as well, are pieces of information about events or results of measurements at a specific time instant, for instance, humidity measured by a sensor near a condenser. Based on the digital twin data platform developed by the authors, we demonstrate the feasibility of adding semantics to the real-time data produced by sensors. Of course, the integration of customised fault tags with real-time data is applicable in other data platforms with alike data architectures.

3.3.1. Digital twin data platform

The Adaptive City Platform (ACP)[4] [56] is adopted as the digital twin data platform in this study, responsible for integrating real-time data with corresponding fault tags. As highlighted in the landmark paper about the semantic web [14], developments will usher in significant new functionality as machines become much better able to process and 'understand' the data that they merely display at present. In the case of real-time data mostly coming from IoT sensors, the data messages usually contain very limited semantics (e.g., sensor ID) in addition to the raw data. Therefore, adding extra semantics (e.g., fault tags) to stream data without breaking the constraints on resource usage becomes increasingly important, as appended data become independent and self-contained. The ACP is designed as a data lake where metadata and stream data are integrated in this platform following an Extract, Transform, Load (ETL) process, and stored for later use. Subsequently, 'data pipelines' are defined to filter (according to the annotated fault tag) and expose extant data as required by functionalities, such as FDD.

According to the comparison of different formats in [57], in the IoT domain, the JavaScript Object Notation (JSON) format with entity-centric structure show advantages in terms of expressivity, query execution time and resource consumption when compared with a triplets-centric structure like RDF (e.g., Brick model in turtle format). The ACP is engineered towards minimising the end-to-end latency for real-time data in the JSON format. The latency between a data entry (i.e., when it is ingested) and exit (i.e., when it is available for use) averages a few milliseconds. To ensure better compatibility and fast querying, the ACP data modelling strategy opts for transforming both the brick model, including metadata and defined fault tags, as well as real-time data, to a more flexible 'crate model' in the JSON format. The transformation process of Brick model (Brick2ACP) is elaborated in Merino et al. [58]. A crate is an entity with its own attributes plus zero or more parents. All crates together with their connections to parents form a hierarchical structure, although it seems that every crate sits at the same level. Every crate is uniquely identified through an indexed key for quick access and query. Listing 1 shows an example of the transformed crate data in the JSON format.

Listing 1: Example of transformed crate data in the JSON format

```
"PointA": {
  "point_name": "PointA",
  "type": "https://brickschema.org/schema/Brick#
     Temperature_Sensor",
  "parents": [
    { "parent_id": "EquipmentB",
```

[4] https://pages.cdbb.uk/projects/

```
      "type": "https://brickschema.org/schema/
          Brick#hasPoint",
      "parent_type": "equipment"
    } ]
},
"ZoneC": {
  "location_name": "ZoneC",
  "type": "https://brickschema.org/schema/Brick#
      HVAC_Zone",
  "parents": [
    { "parent_id": "EquipmentB",
      "type": "https://brickschema.org/schema/
          Brick#feeds",
      "parent_type": "equipment"
    } ]
},
"EquipmentB": {
  "equipment_name": "EquipmentB",
  "type": "https://brickschema.org/schema/Brick#
      VAV",
  "parents": []
}
```

Metadata, particularly the fault tags defined in Section 3.2, can be transformed to be in line with the ACP data modelling strategy. Sharing the same format as real-time data, timestamps can be added easily to track the changes of metadata and fault tags. More importantly, the crate model stays in memory, which accelerate the annotation of real-time data messages with metadata and fault tags.

3.3.2. Integration of knowledge tags with real-time data

`Point` in Brick schema represents devices attached to a location or equipment that generate periodic data (e.g., every minute) or event data (e.g., alerts). Real-time data from points (sensors in this context) is always timestamped to record the time instant when that measurement or event happened. Conceptually, real-time data belongs to the `equipment` or `location` that the sensor monitors, and likewise the fault tags that the sensor is affiliated with. Once the new sensor reading arrives through the MQTT protocol (Message Queuing Telemetry Transport, a standard messaging protocol for the IoT), it comes with point/sensor identifier, timestamp and reading (value). According to the sensor identifier, the metadata of corresponding equipment or location and fault tags should be integrated into the sensor reading in the crate format. Listings 2 and 3 show the format of original temperature sensor readings transmitted through MQTT and the appended reading in the ACP.

Listing 2: Example of temperature sensor reading

```
{
  "Point_id": "Amb_temp_203",
  "Temperature": 21.59,
  "timestamp": 133864886.421
}
```

Listing 3: Annotated temperature sensor reading

```
{
  "Point_id": "Amb_temp_203",
  "Temperature": 21.59,
  "timestamp": 133864886.421,
  "Room203": {
    "location_name": "Room203",
    "type": "https://brickschema.org/schema/
        Brick#Room"
  },
  "FaultA": {
    "tag_name": "FaultA",
    "type": "https://brickschema.org/schema/
        Brick#Tag"
  }
}
```

As shown in Fig. 4, data pipelines can be customised in the digital twin data platform to search or subscribe to the real-time data containing a specific fault tag. The Brick model is transformed into the crate data model and the real-time data stream is annotated with the metadata of corresponding equipment, location and fault tags. Finally, the real-time data stored as a sequence of JSON objects can be filtered, and the data containing the targeted fault tag is then exposed and published to the FDD functionality, which drives the FDD functionality using the Kullback–Leibler divergence.

3.4. Revealing faults through symbolic time-series representations

SAX is also a promising technology available to reduce the dimensionality of the time-series, while keeping the essential information. Note that a and w/p determine the information loss of the transformation and furthermore the computational and memory savings of the overall FDD process. The window size w is usually configured around one to several hours, acknowledging that hour-long data is sufficient to expose most common faults. The optimal values for a and p are determined by trial and error. Generally, a and p values resulting in higher χ^2 scores should be adopted, considering that more distinguishing symbolic sequences between normal and faulty conditions can be acquired under these values. By transforming time-series into a symbolic representation, the lower-dimensional symbolic sequence can be effectively coupled with the Kullback–Leibler divergence based Goodness-of-Fit test to reveal the occurrence of specific faults [59]. The Kullback–Leibler divergence is a type of statistical distance measuring the difference between two discrete random variables. Let Q and $\hat{Q}$ denote two discrete random variables, and the Kullback–Leibler divergence of Q from $\hat{Q}$ is defined as follows:

$$D_{KL}(\hat{Q}\|Q) = \sum_{x} \hat{Q}(x) \log \frac{\hat{Q}(x)}{Q(x)} \tag{2}$$

Here, the Kullback–Leibler divergence based goodness-of-fit test is adopted to classify the sliding time-series as faulty or normal by tracking the time-evolving distribution of the generated codewords. Assuming the sliding time-series from the tagged sensor dimensions is transformed into symbolic sequences of codewords, and in this case, Q and $\hat{Q}$ represent the probability histograms of these codewords generated by normal and observed data respectively. As shown in Eq. (3), the goodness-of-test test is performed by comparing $2N$ times the Kullback–Leibler divergence (N is the total number of observed codewords) against the threshold determined by the compositional inverse of the Cumulative Distribution Function (CDF) of the chi-squared distribution, denoted by $F^{-1}_{\chi^2}(\gamma)$. The fault alarm is triggered if the condition in Eq. (3) holds, where γ is typically set equal to 0.05 or 0.01. The condition indicates that the probability distribution $\hat{Q}$ of the alphabet symbols generated from observed data is considered to be asymptotically dissimilar from that of the symbols from normal data, labelled as Q.

$$2N \cdot D_{KL}(\hat{Q}\|Q) \geq F^{-1}_{\chi^2}(\gamma) \tag{3}$$

4. Case study

The proposed digital twin enabled fault detection and diagnosis functionality for building HVAC systems is tested using the experimental data from a $300m^2$ research facility in the Oak Ridge National Laboratory (ORNL) [60]. The facility is reserved for experiments, and the internal loads are designed to emulate the working condition of an office building. During the experiment, this facility is conditioned with a single packaged rooftop unit (RTU), connecting to multi-zone variable air volume (VAV) terminal systems with electric resistance reheat. The outdoor air intake of the RTU is blocked throughout the

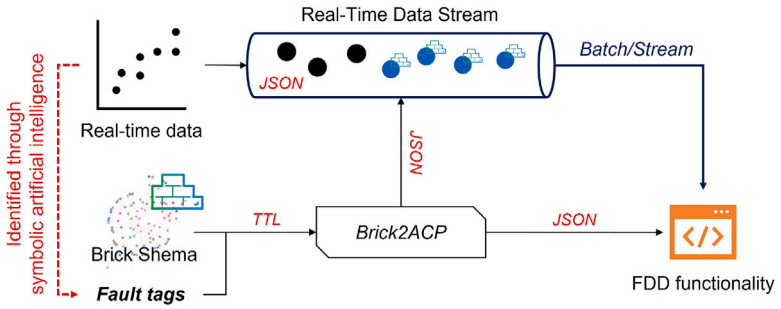

Fig. 4. Data pipelines to support FFD functionality.

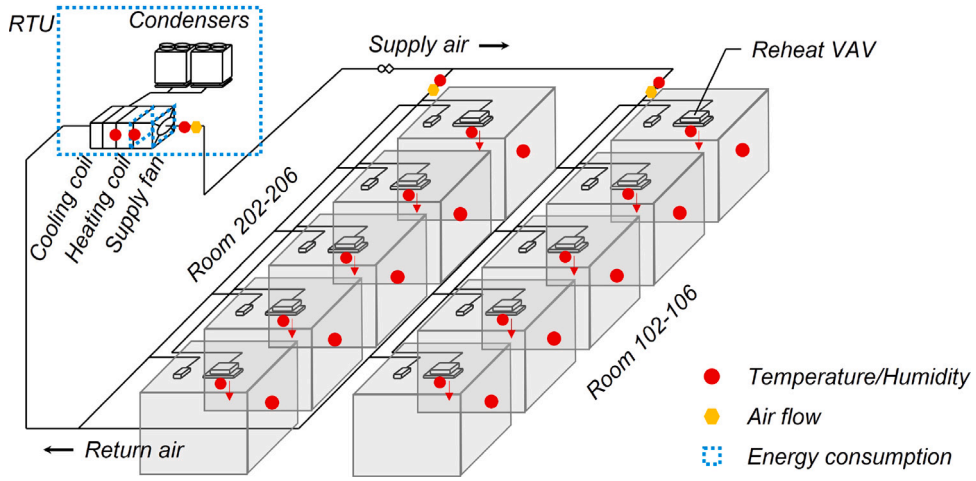

Fig. 5. Schematic of the HVAC system in the facility.

Table 1
Deployed sensors used for the FDD of HVAC system within the facility.

Measuring object	Sensor type	Description
RTU	Air temperature	Measured RTU supply and return air temperature
	Airflow rate	Measured RTU volumetric airflow rate
	Chilled water temperature and pressure	Measured discharge/suction water temperature and pressure to/from the cooling coil
	Condenser water temperature and pressure	Measured condenser outlet refrigerant temperature and pressure
	Energy consumption	Measured RTU electricity and gas consumption
VAV	Air temperature	VAV box discharge air temperature
Zone	Air temperature and humidity	Measured ambient temperature and relative humidity of 10 zones
Lighting system	Energy consumption	Total electricity consumption of lighting system for all 10 zones

experiments, indicating that only indoor air recirculates. Functionally, the RTU and VAVs provide heating and cooling to this experimental building, serving 2 core zones and 8 perimeter zones respectively. Fig. 5 presents the schematic diagram for the air conditioning system of the facility.

The experimental facility is scheduled to operate automatically following the designed occupied and unoccupied modes. The occupied mode starts at 7:00 am and ends at 10:00 pm, while the unoccupied mode lasts for the rest of the day. The cooling coil valve and heating coil valve within the RTU are modulated to maintain the supply air temperature at 55 °F year-round. During the occupied time period, the zone air temperature heating setpoint is 69.8 °F, while the setpoint is reduced to 60 °F during the unoccupied hours.

Various sensors are deployed within the facility to monitor the working condition of the RTU and VAVs with a sampling interval of one minute, including temperature, humidity, pressure, airflow and energy consumption (i.e., electricity, gas) sensors. Table 1 lists the sensors deployed within the facility to realise the fault detection and diagnosis, and the critical sensors have been marked on Fig. 5.

Instead of focusing on the HVAC and lighting schedule setback faults caused by the disordered control sequences, this study targets five types of functional faults that substantially change the behaviour of the HVAC system, including excessive infiltration, thermostat measurement positive and negative biases of core and perimeter zones (+4 °F and −4 °F for core zone 103 and perimeter zone 205). These faults were artificially created and imposed onto the facility: the excessive infiltration is realised by opening windows to achieve the target infiltration rate, and thermostat measurement positive or negative bias is realised by adjusting the temperature setpoint in the opposite direction.

The BoW-based feature extraction and selection method is used to pick up the most distinguishing sensor sets corresponding to each of these five faults. To address missing sensor data values, the labelled data under normal and faulty conditions is first down-sampled to five minutes. Following the data processing procedure shown in Fig. 3, the down-sampled data is sliced in an overlapping manner using a sliding window with size of 12 ($w = 12$, a window of one hour). Data from 4 labelled normal days and 1 faulty day is used to identify the sensitive sensory dimensions for each type of fault. For data in each labelled day, it is converted to 277 codewords of 4 letters from a dictionary of 4 alphabets ($p = 4$, $a = 4$). A χ^2 test is conducted to find the most distinguishing codewords that differentiate normal and five faulty conditions. The sensory dimensions that generate corresponding

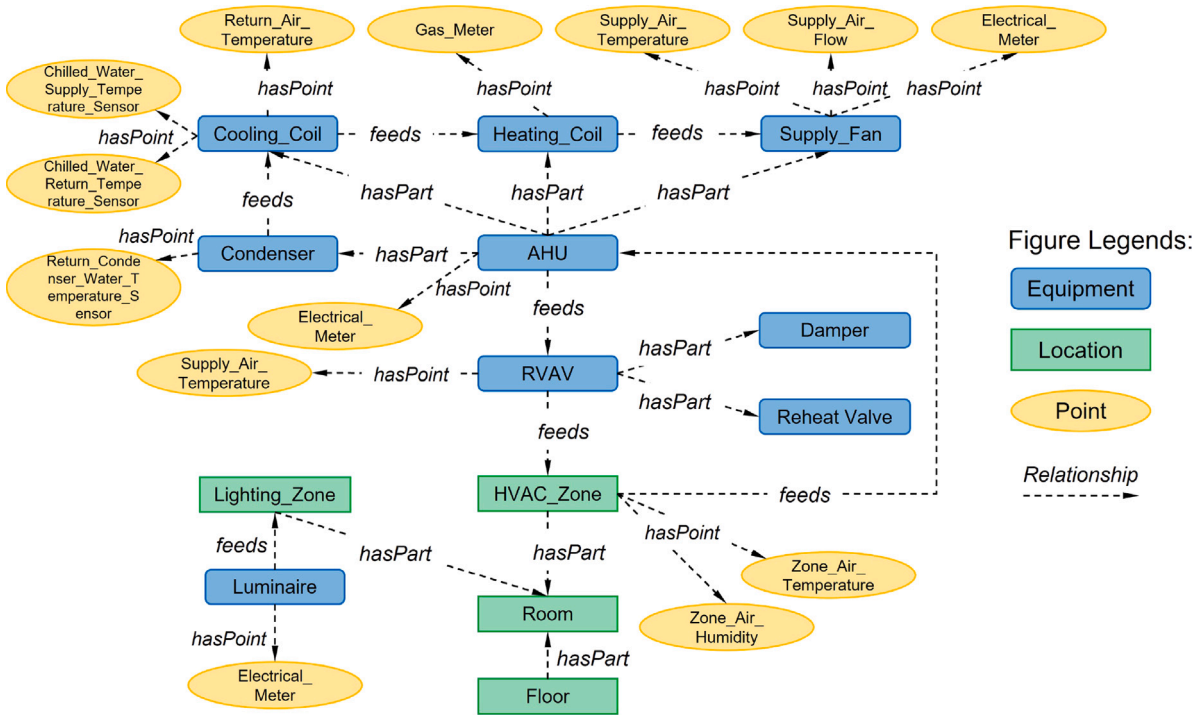

Fig. 6. Brick classes and relationships for the experimental building.

Table 2
Selected sensor sets for each fault type.

	Fault type	Selected sensors
a	Excessive infiltration	Discharge temperature to the cooling coil VAV discharge temperature for zone 202
b	Bias of +4 °F at zone 103	Discharge temperature to the cooling coil VAV discharge temperature for zone 105 VAV discharge temperature for zone 104 Ambient temperature of zone 103
c	Bias of −4 °F at zone 103	Discharge temperature to the cooling coil VAV discharge temperature for zone 106 VAV discharge temperature for zone 105 Ambient temperature of zone 103
d	Bias of +4 °F at zone 205	RTU supply air temperature VAV discharge temperature for zone 202 VAV discharge temperature for zone 204 VAV discharge temperature for zone 205 Ambient temperature for zone 205
e	Bias of −4 °F at zone 205	VAV discharge temperature for zone 204 VAV discharge temperature for zone 205 VAV discharge temperature for zone 206 Ambient temperature for zone 205

codewords with χ^2 score greater than the threshold would be selected. The threshold of 7.879 is selected with the significant level $\alpha = 0.005$. Table 2 presents the selected sensor sets for the FDD of these five fault types.

As shown in Table 2, only 2 to 5 sensors out of the total 51 sensors are seen as informative for the detection and diagnosis of each typical fault respectively. The selection of sensors does follow certain physical knowledge and common sense. For example, the excessive infiltration caused by inappropriately opening windows would directly influence the return air temperature (no return air temperature sensor deployed) and subsequently the discharge temperature of the cooling coil before the air is heated in winter. Besides, the bias of the zone thermostat measurement would be easily noticed from the ambient temperature of the zone. Of course, the correlation between the fault and sensor data and the strength of that correlation can only be seen from data. To preserve this ad-hoc knowledge in the digital twin ecosystem, fault tags need to be supplemented in the corresponding semantic model. Fig. 6 presents the Brick model of the HVAC system in the facility defined using the Brick ontology. The physical and logical entities in the facility, such as the zones, RTU, VAVs, their components and deployed sensors are explicitly defined, and the relationships (e.g., *hasPart*, *feeds*, *hasPoint*) are modelled to capture the connections between these entities. Five tags (i.e., fault_a, fault_b, fault_c, fault_d, fault_e) are added in the Brick model for the five fault types listed in Table 2 and label the sensors that are informative to each fault (i.e., points *hasTag* fault_x).

Leveraging the fault tags, sensitive sensory dimensions to each fault are uniquely labelled. Furthermore, the digital twin data platform annotates the minute-level data stream with metadata including the defined fault tags. As an example, Fig. 7 illustrates a part of the HVAC system of this facility, which contains the five sensors informative to fault_d and as well as the equipment and locations associated with these sensors. The detection of fault_d mainly relies on the following five temperature sensors:

- VAVT_202: discharge temperature of VAV in zone 202,
- VAVT_204: discharge temperature of VAV in zone 204,
- VAVT_205: discharge temperature of VAV in zone 205,
- ZoneT_205: the ambient temperature in zone 205, and
- RValve_205: the RTU supply air temperature.

To guarantee the real-time character of the data integration, the real-time data stream is converted into the JSON format. The sliced Brick model (see Fig. 7) can be transformed into crates, within which the *hasPoint* and *hasTag* connections in the Brick model are highlighted. These connections contain all metadata needed to annotate real-time data streams. Using the fault_d as an example, real-time readings from the five tagged sensors are annotated with the associated equipment, location and particularly the fault tag (fault_d) to create a self-contained message feeding into the FDD functionality (see Fig. 8). Specifically, for the sensor readings with the point identifier of ZoneT_205, the location of the sensor (Zone_205) and the associated fault tag (fault_d) are appended and annotated to the JSON object.

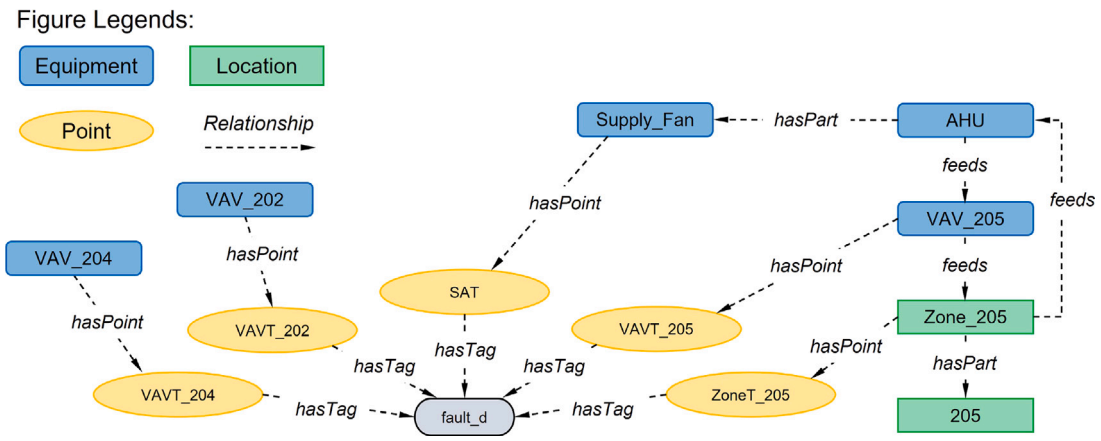

Fig. 7. Part of the HVAC system in the facility associated with fault_d.

Fig. 8. Data tagging of readings with BrickSchema metadata.

As a result, the identified sensory dimensions of massive real-time data are filtered to efficiently feed the FDD functionality through the semantically defined fault tags. That is to say, the data stream annotated with the specific fault tag is filtered and exposed through the data pipelines to feed into the FDD functionality. For the implementation, the real-time sensor data preprocessed in the digital twin data platform are fetched using WebSocket connections [56]. Uniquely, the FDD in this paper is conducted in a real-time manner, by building a probability model of normal data [61]. In this case, the probability histogram Q of codewords transformed from the 4 day normal operational data is computed offline and remains constant during the entire FDD process. The faulty data stream is artificially generated from the experimental facility for each fault type. To showcase the use of the one-sided goodness-of-fit test on revealing faulty scenarios, the fraction of times the codewords transformed from the reduced-dimensional faulty data stream appear in the past few windows (i.e., $\hat{Q}$) is compared against the probability histogram Q. Based on the Goodness-of-Fit test with Kullback–Leibler divergence, the true positive rate is calculated for the five fault types respectively. The results are given in Table 3, indicating that the Goodness-of-Fit test can largely detect the target five faults based on the data provided. It is also worth mentioning that the proposed digital twin analytical framework is compatible with other FDD algorithms as well, which identify the faults based on diverse machine learning techniques (e.g., Artificial Neural Networks) using the filtered data streams as input through the same data pipeline.

Table 3
FDD performance results in terms of true positive rate.

	fault_a	fault_b	fault_c	fault_d	fault_e
True positive	63.8%	61.4%	53.9%	68.7%	70.2%

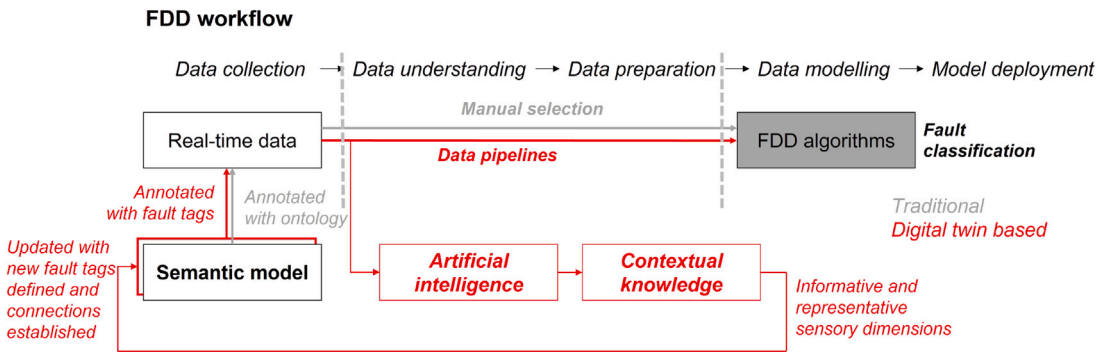

Fig. 9. Improved FDD workflow based on the digital twin analytical framework.

5. Discussion

As shown in Fig. 9, Boi-Ukeme et al. [62] summarised the typical FDD workflow, including data collection from the building systems of interest, data understanding, data preparation and analysis, data modelling, and deployment of the models for FDD. Most of the FDD research focuses on data modelling and its application in fault classification using supervised or unsupervised FDD algorithms. However, the data understanding, preparation and analysis (e.g., exploratory analysis and determining appropriate input for the algorithm), which often require human intervention, present a significant obstacle to the automation of FDD. To reduce human intervention and automate the FDD process, it is necessary to ensure that ad-hoc knowledge learned through statistical and symbolic artificial intelligence techniques from real-time data becomes comprehensible by computers, on top of the existing ontological knowledge. The contextual/temporal knowledge in the FDD process refers to the recognised sensor dimensions dedicated to the detection of specific faults. To preserve the new knowledge for better understanding, this study extends the Brick ontology with fault tags, which support labelling the selected sensors with machine-readable fault tags. By annotating the real-time sensor data with fault tags in the Brick model, the data pipelines prepare the real-time data targeting specific faults to serve the FDD functionality. This is a preliminary trial in realising building intelligence, considering that a massive amount of data from heterogeneous sources must be filtered and prepared before feeding into the corresponding functionalities. It contributes to the dimensionality reduction of the original problem and preparing a reasonable dataset following the 'divide-and-conquer' strategy.

The proposed digital twin enabled FDD process contributes to the automation in monitoring building critical assets. Instead of performing manual inspections according to a predefined schedule, the FDD automatically informs the facility managers of the emergence of faults. Meanwhile, the fault watch list is defined by the facility management professionals that comes with historical work order lists for different faults. The historical work orders can be used to suggest corrective actions based on the past maintenance procedures, implemented to maintain or repair corresponding assets that lead to the emerging fault. The proposed FDD process can be integrated with existing asset management databases, for instance the Computer Aided Facility Management (CAFM) software [63], where the work order data can be managed. This is the first step towards the automation of Operations Maintenance and Repair (OM & R). It would be of interest here to expand the static building system representation with the plethora of building information available in a Linked Building Data (LBD) cloud to enable even more holistic analyses for asset management. Automating the FDD process also empowers other intelligent building functionalities, for example the detected fault history can be used to train machine learning models to predict the occurrence of future faults of the same type. This would support the further automation of the asset management processes.

6. Conclusions

Abundant data, particularly real-time data, is generated in dynamic building systems. These pieces of data, carrying a bewildering amount of information, need to be selected and filtered to drive corresponding intelligent building functionalities, otherwise, the irrelevant data dimensions would mask and flush the most informative data. In the case of FDD, the Bag-of-Words based (i.e., SAX) feature extraction and selection method, as a typical symbolic artificial intelligence technique, is introduced to boost computational efficiency. It tries to extract the temporal characteristics of historical time-series and identify the distinguishing sensory dimensions from the labelled normal and faulty data, which deepens the understanding of all the concerned faults. The 'fault tags' are defined accordingly in the Brick model to label these relevant sensor entities, preserving the learned contextual knowledge/understanding in a machine-readable way. The digital twin data platform is adopted to integrate the defined knowledge tags with real-time data. By annotating the data stream with auxiliary fault tags, low-latency high-bandwidth real-time data streams appended with the specified fault tag can be automatically extracted to feed the FDD functionality. This informs the way to enable dynamic asset management functionalities through digital twinning. The divide-and-conquer strategy used here contributes to the real-time character and the computational burden reduction for delivering building intelligent functionalities.

To the future of digital twins, the proposed framework allows to overcome the problem of 'big' data (high volume, high variety, but not high velocity) and filter the large amount of data coming from the digital twin ecosystem. Thanks to the metadata tagging and annotating approaches, the developed methodologies can be replicated and allow to take good advantage of ontological knowledge and ad-hoc knowledge from artificial intelligence techniques to support intelligent building functionalities. The expressiveness and richness of semantics are leveraged for developing complex functionalities, allowing to better represent comprehensive understandings about the building systems and drive the real-time management of the physical assets. However, additional research efforts are needed to achieve further support and automation in the decision-making process (e.g., through the integration with CAFM and other organisational systems). The data integration approach is based on the digital twin data platform, prioritising the real-time character of data and treating the pseudo-static information as metadata appended to real-time data streams. It makes this approach flexible and very effective for the development of dynamic intelligent building functionalities (e.g., asset management applications).

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.

Data availability

Data will be made available on request.

Acknowledgements

This research forms part of the Centre for Digital Built Britain's (CDBB) work at the University of Cambridge within the Construction Innovation Hub (CIH). The Construction Innovation Hub is funded by UK Research and Innovation, UK through the Industrial Strategy Fund. Furthermore, the funding support by the Dutch Netherlands Enterprise Agency for the Brains4Buildings project is acknowledged as well to make part of this work possible.

References

[1] N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, A.M. Tsang, P. Switzer, J.V. Behar, S.C. Hern, W.H. Engelmann, The national human activity pattern survey (NHAPS): A resource for assessing exposure to environmental pollutants, J. Expo. Sci. Environ. Epidemiol. 11 (3) (2001) 231–252, http://dx.doi.org/10.1038/sj.jea.7500165.

[2] Brits spend 90% of their time indoors, 2018. https://www.opinium.com/brits-spend-90-of-their-time-indoors/. (Accessed 19 October 2022).

[3] J. Wong, H. Li, J. Lai, Evaluating the system intelligence of the intelligent building systems: Part 1: Development of key intelligent indicators and conceptual analytical framework, Autom. Constr. 17 (3) (2008) 284–302, http://dx.doi.org/10.1016/j.autcon.2007.06.002.

[4] A.P. Plageras, K.E. Psannis, C. Stergiou, H. Wang, B.B. Gupta, Efficient IoT-based sensor BIG data collection–processing and analysis in smart buildings, Future Gener. Comput. Syst. 82 (2018) 349–357, http://dx.doi.org/10.1016/j.future.2017.09.082.

[5] M. Grieves, J. Vickers, Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems, in: Transdisciplinary Perspectives on Complex Systems, Springer, 2017, pp. 85–113, http://dx.doi.org/10.1007/978-3-319-38756-7_4.

[6] C. Boje, A. Guerriero, S. Kubicki, Y. Rezgui, Towards a semantic construction digital twin: Directions for future research, Autom. Constr. 114 (2020) 103179, http://dx.doi.org/10.1016/j.autcon.2020.103179.

[7] E. Seghezzi, M. Locatelli, L. Pellegrini, G. Pattini, G.M. Di Giuda, L.C. Tagliabue, G. Boella, Towards an occupancy-oriented digital twin for facility management: Test campaign and sensors assessment, Appl. Sci. 11 (7) (2021) 3108, http://dx.doi.org/10.3390/app11073108.

[8] Q. Lu, X. Xie, A.K. Parlikad, J.M. Schooling, Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance, Autom. Constr. 118 (2020) 103277, http://dx.doi.org/10.1016/j.autcon.2020.103277.

[9] M.A. Jafari, E. Zaidan, A. Ghofrani, K. Mahani, F. Farzan, Improving building energy footprint and asset performance using digital twin technology, in: 4th IFAC Workshop on Advanced Maintenance Engineering, Services and Technologies, Vol. 53, (3) 2020, pp. 386–391, http://dx.doi.org/10.1016/j.ifacol.2020.11.062.

[10] E. O'Dwyer, I. Pan, R. Charlesworth, S. Butler, N. Shah, Integration of an energy management tool and digital twin for coordination and control of multi-vector smart energy systems, Sustainable Cities Soc. 62 (2020) 102412, http://dx.doi.org/10.1016/j.scs.2020.102412.

[11] E.A. Pärn, D.J. Edwards, M.C.P. Sing, The building information modelling trajectory in facilities management: A review, Autom. Constr. 75 (2017) 45–55, http://dx.doi.org/10.1016/j.autcon.2016.12.003.

[12] E.A. Rogers, R.N. Elliott, S. Kwatra, D. Trombley, V. Nadadur, Intelligent efficiency: Opportunities, barriers, and solutions, in: Automated Diagnostics and Analytics for Buildings, River Publishers, 2021, pp. 35–71, ISBN: 9781003151906.

[13] M.S. Mirnaghi, F. Haghighat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review, Energy Build. (2020) 110492, http://dx.doi.org/10.1016/j.enbuild.2020.110492.

[14] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Sci. Am. 284 (5) (2001) 34–43, URL https://www.jstor.org/stable/26059207#metadata_info_tab_contents. (Accessed 19 October 2022).

[15] D. Jung, C. Sundström, A combined data-driven and model-based residual selection algorithm for fault detection and isolation, IEEE Trans. Control Syst. Technol. 27 (2) (2017) 616–630, http://dx.doi.org/10.1109/TCST.2017.2773514.

[16] R. Liu, D.F. Gillies, Overfitting in linear feature extraction for classification of high-dimensional image data, Pattern Recognit. 53 (2016) 73–86, http://dx.doi.org/10.1016/j.patcog.2015.11.015.

[17] S. Guo, W. Guo, Process monitoring and fault prediction in multivariate time series using bag-of-words, IEEE Trans. Autom. Sci. Eng. (2020) 230–242, http://dx.doi.org/10.1109/TASE.2020.3026065.

[18] A. Bolton, L. Butler, I. Dabson, M. Enzer, M. Evans, T. Fenemore, F. Harradence, E. Keaney, A. Kemp, A. Luck, et al., Gemini principles, 2018, http://dx.doi.org/10.17863/CAM.32260.

[19] J. Schooling, M. Enzer, D.G. Broo, Flourishing systems: Re-envisioning infrastructure as a platform for human flourishing, Proc. Inst. Civ. Eng. Smart Infrastruct. Constr. 173 (1) (2021) 166–174, http://dx.doi.org/10.1680/jsmic.20.00023.

[20] P. Pauwels, S. Zhang, Y.-C. Lee, Semantic web technologies in AEC industry: A literature overview, Autom. Constr. 73 (2017) 145–165, http://dx.doi.org/10.1016/j.autcon.2016.10.003.

[21] C. Eastman, The use of computers instead of drawings in building design, AIA J. 63 (3) (1975) 46–50.

[22] J.X. Parreira, D. Dhungana, G. Engelbrecht, The role of RDF stream processing in an smart city ICT infrastructure-the aspern smart city use case, in: European Semantic Web Conference, Springer, 2015, pp. 343–352, http://dx.doi.org/10.1007/978-3-319-25639-9_47.

[23] R. Tommasini, E. Della Valle, A. Mauri, M. Brambilla, RSPLab: RDF stream processing benchmarking made easy, in: International Semantic Web Conference, Springer, 2017, pp. 202–209, http://dx.doi.org/10.1007/978-3-319-68204-4_21.

[24] X. Xie, Q. Lu, M. Herrera, Q. Yu, A.K. Parlikad, J.M. Schooling, Does historical data still count? Exploring the applicability of smart building applications in the post-pandemic period, Sustainable Cities Soc. 69 (2021) 102804, http://dx.doi.org/10.1016/j.scs.2021.102804.

[25] S. Mishra, A. Glaws, D. Cutler, S. Frank, M. Azam, F. Mohammadi, J.-S. Venne, Unified architecture for data-driven metadata tagging of building automation systems, Autom. Constr. 120 (2020) 103411, http://dx.doi.org/10.1016/j.autcon.2020.103411.

[26] B. Zhong, H. Wu, H. Li, S. Sepasgozar, H. Luo, L. He, A scientometric analysis and critical review of construction related ontology research, Autom. Constr. 101 (2019) 17–31, http://dx.doi.org/10.1016/j.autcon.2018.12.013.

[27] P. Pauwels, A. Costin, M.H. Rasmussen, Knowledge graphs and linked data for the built environment, in: Industry 4.0 for the Built Environment, Springer, 2022, pp. 157–183, http://dx.doi.org/10.1007/978-3-030-82430-3_7.

[28] W. Terkaj, A. Šojić, Ontology-based representation of IFC EXPRESS rules: An enhancement of the ifcOWL ontology, Autom. Constr. 57 (2015) 188–201, http://dx.doi.org/10.1016/j.autcon.2015.04.010.

[29] P. Pauwels, W. Terkaj, EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology, Autom. Constr. 63 (2016) 100–133, http://dx.doi.org/10.1016/j.autcon.2015.12.003.

[30] M.H. Rasmussen, P. Pauwels, C.A. Hviid, J. Karlshoj, Proposing a central AEC ontology that allows for domain specific extensions, in: Joint Conference on Computing in Construction, Vol. 1, 2017, pp. 237–244, http://dx.doi.org/10.24928/JC3-2017/0153.

[31] M.H. Rasmussen, M. Lefrançois, G.F. Schneider, P. Pauwels, BOT: The building topology ontology of the W3C linked building data group, Semantic Web 12 (1) (2021) 143–161, http://dx.doi.org/10.3233/SW-200385.

[32] M. Poveda-Villalón, R. Garcıa-Castro, Extending the SAREF ontology for building devices and topology, in: Proceedings of the 6th Linked Data in Architecture and Construction Workshop, LDAC 2018, Vol. CEUR-WS, Vol. 2159, 2018, pp. 16–23, URL http://ceur-ws.org/Vol-2159/02paper.pdf. (Accessed 19 October 2022).

[33] M. Rasmussen, M. Lefrançois, M. Bonduel, C. Hviid, J. Karlshøj, OPM: An ontology for describing properties that evolve over time, in: 6th Linked Data in Architecture and Construction Workshop, 2018, pp. 24–33, URL http://ceur-ws.org/Vol-2159/03paper.pdf. (Accessed 19 October 2022).

[34] A. Wagner, U. Rüppel, BPO: The building product ontology for assembled products, in: Proceedings of the 7th Linked Data in Architecture and Construction Workshop, LDAC 2019', Lisbon, Portugal, 2019, pp. 106–119, URL http://ceur-ws.org/Vol-2389/08paper.pdf. (Accessed 19 October 2022).

[35] A. Wagner, W. Sprenger, C. Maurer, T.E. Kuhn, U. Rüppel, Building product ontology: Core ontology for linked building product data, Autom. Constr. 133 (2022) 103927, http://dx.doi.org/10.1016/j.autcon.2021.103927.

[36] Brick: A uniform metadata schema for buildings, https://brickschema.org/. (Accessed 19 October 2022).

[37] B. Balaji, A. Bhattacharya, G. Fierro, J. Gao, J. Gluck, D. Hong, A. Johansen, J. Koh, J. Ploennigs, Y. Agarwal, et al., Brick: Metadata schema for portable smart building applications, Appl. Energy 226 (2018) 1273–1292, http://dx.doi.org/10.1016/j.apenergy.2018.02.091.

[38] G. Fierro, J. Koh, Y. Agarwal, R.K. Gupta, D.E. Culler, Beyond a house of sticks: Formalizing metadata tags with brick, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 125–134, http://dx.doi.org/10.1145/3360322.3360862.

[39] V. Kukkonen, A. Kücükavci, M. Seidenschnur, M.H. Rasmussen, K.M. Smith, C.A. Hviid, An ontology to support flow system descriptions from design to operation of buildings, Autom. Constr. 134 (2022) 104067, http://dx.doi.org/10.1016/j.autcon.2021.104067.

[40] X. Xie, N. Moretti, J. Merino, J.Y. Chang, A.K. Parlikad, Ontology-based spatial and system hierarchies federation for fine-grained building energy analysis, in: Proc. of the Conference CIB W78, 2021, pp. 368–377, URL http://itc.scix.net/paper/w78-2021-paper-037. (Accessed 19 October 2022).

[41] G. Fierro, J. Koh, S. Nagare, X. Zang, Y. Agarwal, R.K. Gupta, D.E. Culler, Formalizing tag-based metadata with the brick ontology, Front. Built Environ. (2020) 152, http://dx.doi.org/10.3389/fbuil.2020.558034.
[42] 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, http://dx.doi.org/10.1016/j.ijrefrig.2006.12.012.
[43] B. Sun, P.B. Luh, Q.-S. Jia, Z. O'Neill, F. Song, Building energy doctors: An SPC and Kalman filter-based method for system-level fault detection in HVAC systems, IEEE Trans. Autom. Sci. Eng. 11 (1) (2013) 215–229, http://dx.doi.org/10.1109/TASE.2012.2226155.
[44] Y. Zhao, J. Wen, F. Xiao, X. Yang, S. Wang, Diagnostic Bayesian networks for diagnosing air handling units faults—Part I: Faults in dampers, fans, filters and sensors, Appl. Therm. Eng. 111 (2017) 1272–1286, http://dx.doi.org/10.1016/j.applthermaleng.2015.09.121.
[45] Y. Zhao, J. Wen, S. Wang, Diagnostic Bayesian networks for diagnosing air handling units faults—Part II: Faults in coils and sensors, Appl. Therm. Eng. 90 (2015) 145–157, http://dx.doi.org/10.1016/j.applthermaleng.2015.07.001.
[46] Z.J. Yu, F. Haghighat, B.C. Fung, L. Zhou, A novel methodology for knowledge discovery through mining associations between building operational data, Energy Build. 47 (2012) 430–440, http://dx.doi.org/10.1016/j.enbuild.2011.12.018.
[47] E. Petrova, P. Pauwels, K. Svidt, R.L. Jensen, In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data, in: Advances in Informatics and Computing in Civil and Construction Engineering, Springer, 2019, pp. 19–26, http://dx.doi.org/10.1007/978-3-030-00220-6_3.
[48] K. Yan, W. Shen, T. Mulumba, A. Afshari, ARX model based fault detection and diagnosis for chillers using support vector machines, Energy Build. 81 (2014) 287–295, http://dx.doi.org/10.1016/j.enbuild.2014.05.049.
[49] T. Mulumba, A. Afshari, K. Yan, W. Shen, L.K. Norford, Robust model-based fault diagnosis for air handling units, Energy Build. 86 (2015) 698–707, http://dx.doi.org/10.1016/j.enbuild.2014.10.069.
[50] D. Li, Y. Zhou, G. Hu, C.J. Spanos, Optimal sensor configuration and feature selection for AHU fault detection and diagnosis, IEEE Trans. Ind. Inform. 13 (3) (2016) 1369–1380, http://dx.doi.org/10.1109/TII.2016.2644669.
[51] L. Zhang, S. Frank, J. Kim, X. Jin, M. Leach, A systematic feature extraction and selection framework for data-driven whole-building automated fault detection and diagnostics in commercial buildings, Build. Environ. 186 (2020) 107338, http://dx.doi.org/10.1016/j.buildenv.2020.107338.
[52] M. Kim, S.H. Yoon, P.A. Domanski, W.V. Payne, Design of a steady-state detector for fault detection and diagnosis of a residential air conditioner, Int. J. Refrig. 31 (5) (2008) 790–799, http://dx.doi.org/10.1016/j.ijrefrig.2007.11.008.
[53] P. Schäfer, U. Leser, Multivariate time series classification with WEASEL+ MUSE, 2017, http://dx.doi.org/10.48550/arXiv.1711.11343.
[54] Y. Chen, G. Lin, E. Crowe, J. Granderson, Development of a unified taxonomy for HVAC system faults, Energies 14 (17) (2021) 5581, http://dx.doi.org/10.3390/en14175581.
[55] V. Rajagopalan, A. Ray, R. Samsi, J. Mayer, Pattern identification in dynamical systems via symbolic time series analysis, Pattern Recognit. 40 (11) (2007) 2897–2907, http://dx.doi.org/10.1016/j.patcog.2007.03.007.
[56] J. Brazauskas, R. Verma, V. Safronov, M. Danish, J. Merino, X. Xie, I. Lewis, R. Mortier, Data management for building information modelling in a real-time adaptive city platform, 2021, http://dx.doi.org/10.48550/arXiv.2103.04924.
[57] X. Su, J. Riekki, J.K. Nurminen, J. Nieminen, M. Koskimies, Adding semantics to internet of things, Concurr. Comput.: Pract. Exper. 27 (8) (2015) 1844–1860, http://dx.doi.org/10.1002/cpe.3203.
[58] J. Merino, X. Xie, N. Moretti, J.Y. Chang, A. Parlikad, Data integration for digital twins in the built environment, in: European Conference on Computing in Construction, Rhodes, Greece, 2022, http://dx.doi.org/10.35490/EC3.2022.172.
[59] K. Bountrogiannis, G. Tzagkarakis, P. Tsakalides, Anomaly detection for symbolic time series representations of reduced dimensionality, in: 2020 28th European Signal Processing Conference, EUSIPCO, IEEE, 2021, pp. 2398–2402, http://dx.doi.org/10.23919/Eusipco47968.2020.9287474.
[60] J. Granderson, G. Lin, A. Harding, P. Im, Y. Chen, Building fault detection data to aid diagnostic algorithm creation and performance testing, Sci. Data 7 (1) (2020) 1–14, http://dx.doi.org/10.1038/s41597-020-0398-6.
[61] C. Wang, K. Viswanathan, L. Choudur, V. Talwar, W. Satterfield, K. Schwan, Statistical techniques for online anomaly detection in data centers, in: 12th IFIP/IEEE International Symposium on Integrated Network Management, IM 2011 and Workshops, IEEE, 2011, pp. 385–392, http://dx.doi.org/10.1109/INM.2011.5990537.
[62] J. Boi-Ukeme, G. Wainer, A workflow for data-driven fault detection and diagnosis in buildings, in: 2021 Winter Simulation Conference, WSC, IEEE, 2021, pp. 1–12, http://dx.doi.org/10.1109/WSC52266.2021.9715464.
[63] N. Wills, J. Diaz, Integration of real-time data in BIM enables FM processes, WIT Trans. Built Environ. 169 (2017) 127–133, http://dx.doi.org/10.2495/BIM170121.