Multi-view Data Mining Approach for Behaviour Analysis of Smart Control Valve

Amirmohammad Eghbalian, Shahrooz Abghari, Veselka Boeva
Department of Computer Science, Blekinge Institute of Technology
Karlskrona, Sweden
Email: amirme1992@gmail.com, {sab,vbx}@bth.se

Farhad Basiri
iquest AB
Hägersten, Sweden
Email: farhad.basiri@iquest.se

***Abstract*—In this study, we propose a multi-view data analysis approach that can be used for modelling and monitoring smart control valve system behaviour. The proposed approach consists of four distinctive steps: (i) multi-view interpretation of the available data attributes by separating them into several representations (views), e.g., operational parameters, contextual factors, and performance indicators; (ii) modelling different control valve system operating modes by clustering analyses of the operational data view; (iii) annotating each operating mode (cluster) by using the remaining views (i.e., contextual and system performance data); (iv) context-aware monitoring of the control valve system operating behaviour by applying the built model. In addition, the data points (daily profiles) observed during the monitoring can be annotated by comparing them with the known typical behavioural modes. This information can be further analysed and used for continuous updating and improvement of the model.**

The potential of the proposed approach has been evaluated and demonstrated on real-world sensor data originating from a company in the smart building domain. The obtained results show the robustness of the proposed approach in modelling, analysing, and monitoring the control valve system behaviour.

***Index Terms*—Clustering analysis, Multi-view data mining, Outlier detection, Continuous learning**

I. Introduction

Nowadays, a countless number of devices are equipped with sensors that can communicate together and can be accessed via the Internet. This is referred to as Internet of Things (IoT), which makes our life, city, and premises more modern due to daily advances in different fields like computing, communication, and electronics. Smart environment is a key phrase in IoT. One of the domains covered by IoT is smart buildings. The solutions that have been provided by IoT have a great impact on decreasing energy waste that is caused by sub-optimal asset management and human activities. Some of the examples of IoT automation systems that already exist are SmartThings [1], Vera [2], and openHAB [3].

Heating, ventilation, air conditioning, and refrigeration (HVAC&R) is a system designed to resolve the thermal needs and requirements for different types of buildings such as residential, industrial, and others. The main task of an HVAC&R system is to heat up or cool down the outdoor air according to a desired and required indoor temperature at a building. The important part of the HVAC&R system is its control system which is responsible for regulating the operation and performance of the HVAC&R. In addition, energy management and safety are other capabilities that are expected from modern control systems.

Control valves, as part of any HVAC&R system, are used to modulate the flow and pressure of fluids or gases. The control valves are used as the main final element in the control systems of many equipment. One of the technological advances in smart buildings is the advent of smart control valves. These types of valves are equipped with sensors capable of collecting diagnostic data such as valve position or performance. The sensors collect a large amount of data capturing measurements of different nature. The main challenge about the smart control valves is processing and analysing this large volume of heterogeneous often unlabeled data and extracting useful information about the system's behaviour and performance. Such knowledge can be of great benefit for domain experts in the better understanding and interpreting the system operating modes with respect to different contextual factors, such as outdoor temperature, and identifying deviating behaviours. For instance, the deviating operational behaviour of the system can be an indication of faults, e.g., cavitation or misconfiguration, and in some cases the unsuitable size of the control valve.

In this study, we propose a multi-view data analysis approach that can tackle the above-discussed challenges by proposing a pipeline of techniques for modelling and monitoring smart control valve system behaviour. The proposed approach partitions the available data attributes into several representations (views), e.g., operational parameters, contextual factors, and performance indicators. The motivation is that each view represents part of the relevant information about the system and is worth to be analysed separately in order to better understand the system behaviour and performance. We initially model different operating modes of a control valve system by clustering analysis of the operational data view. Each cluster representative is interpreted as a behavioural signature of typical operating mode (TOM). Further each TOM is annotated by using the remaining views, i.e., the contextual and system performance data. This allows extracting implicit relationships among the different views. For example, some operational behaviours can be grouped together, because they

This work is part of the research project "*Scalable resource-efficient systems for big data analytics*" funded by the Knowledge Foundation (grant: 20140032) in Sweden.

are linked to the one and the same context. Evidently, TOMs concern the internal working behaviour of the system and contextual factors such as outdoor temperature. In addition, each TOM is associated with expected performance. For instance, it would be useful to discover that the same performance may be induced by different TOMs of the system given the same context. Thus the built model can be applied for context-aware monitoring of the control valve system operating behaviour. The approach additionally provides an opportunity for annotation and analysis of the observed data points and continuous updating and refinement of the model when new behavioural modes appeared.

II. Related Work

Fault detection and diagnostics (FDD) is the process of detecting faults and understanding their cause(s) in a physical system [4], [5]. HVAC&R is an example of a physical system in buildings. Automated fault detection and diagnostic (AFDD) equipment are the tools and technologies that can be applied to automate the FDD process.

Active study related to FDD in HVAC&R systems started in the 1980s and since then FDD and data mining techniques have matured considerably [6], [7]. In the later 1980s, AFDD methods on refrigeration based on vapor-compression were published in [8], [9]. In the 1990s, the majority of the applications related to FDD concentrated on vapor-compression devices and air-handling units (AHUs). Generally, these applications were using temperature and/or pressure measurement for general FDD.

From almost 200 published studies related to AFDD for building systems until 2018, around 62% belonged to process history-based (data-driven) AFDD methods, 26% related to qualitative model-based such as rule-based models, and the last 12% focused on quantitative (mathematical) model-based methods. There are two reasons for the popularity of process history-based methods 1) these methods use historical data for creating a model 2) the built model has much lower complexity compared to two other categories. Among the studies that applied process history-based methods, 72% derived from black box models, 12% obtained from gray box models, and 16% acquired from a combination of those two methods. The black box models can further be classified into pattern recognition, statistical, and artificial neural network (ANN) techniques [10].

Ren et al. [11] applied support vector machine (SVM) to classify patterns in a refrigeration system. SVM was used to recognize the best pattern for identifying faults among seven operating patterns (one represents normal state and six represent fault states). Najafi et al. [12] presented an AFDD approach based on the Bayesian network for the air-handling unit diagnosis. The proposed approach was used to analyse and compare the current behavioural patterns of the system with the faulty behavioural patterns, produced by the system faults, to select the most similar patterns that can demonstrate the current behaviour of the system.

Cui and Wang [13] proposed an on-line AFDD technique to demonstrate a centrifugal chiller system health state. In another study, the authors [14] proposed an approach for fault detection in the AHU using autoregressive-moving-average. The method uses a threshold to measure the performance degradation of the system caused by an existing fault to notify when the system needs to be serviced. Armstrong et al. [15] proposed a fifth-order autoregressive model that is able to identify faults such as compressor valve leakage, in the rooftop units by getting one input. Du et al. [16] introduced a black box model-based method using a joint angle analysis that can detect faults in variable-air volume systems.

Kim et al. [17] presented an AFDD method based on black box ANN technique for the air-conditioning system of a residential building. In this AFDD method, probabilities of the normal and faulty states of the system were calculated and based on that the status of the system was notified. Fan et al. [18] presented an AFDD method for AHU that was based on ANN models and wavelet analysis. In this method, a threshold based on the normal operation of the system was selected which was used for identification of the system's faults in comparison to the residuals.

While the majority of studies reviewed in this section focused on the refrigeration system, chiller system, variable-air volume, compressor valve in rooftop units, and air-conditioning systems, we are interested in smart control valves due to their importance for space heating and cooling in the HVAC&R systems. One of the factors that are often neglected in the reviewed works is the limited number of labeled data sets in case of real-world systems such as the smart control valves. Traditional supervised machine learning approaches rely on labeled and observable data sets. However, inefficient labor work of labeling data usually makes this process expensive and useless. Therefore, there is a crucial need to develop learning techniques that are capable of training the model using limited (or none) labeled data to learn the system behaviour.

The above motivated us to propose an unsupervised multi-view approach that is capable of modelling typical operating (behavioral) modes taking into account different characteristics of the studied control valve system, monitoring its behaviour over time, and detecting deviating behaviours. The proposed approach supplies the domain experts with complementary information that facilitates further the understanding and analysis of the system's behaviour and performance. In addition, the built behavioural model is not static, i.e., our approach provides an opportunity for continuous learning and updating of the model when new information becomes available.

III. Methods and Proposed Approach

A. Partitioning Algorithms

The review of the studies related to time-series data analysis over the past decade shows that partitioning algorithms are the ones that are widely used [19]. The fast response and simplicity of these types of algorithms is one of the reasons they are relatively popular. k-means [19], k-medians, and

k-medoids are widely used for analysing time-series data and grouping the data points into k disjoint clusters. The number of clusters (k) is preliminarily determined for the three partitioning algorithms. Each partitioning algorithm initially selects a set of k cluster centers. Each data point is then associated with the nearest cluster, and the cluster centers are recalculated. The process is iterative and it continues until no further reassignments are needed. The difference among the three partitioning algorithms is in how the cluster center is defined. The cluster center in k-means is the mean data vector calculated over all the data objects in the cluster. Instead, the median data vector is used in k-medians. k-medoids is considered as a robust version of the k-means and the cluster center is the most centrally placed point of the cluster.

In this study, k-means is applied for integration analysis of multiple data sets containing measurements of different time-monitored characteristics of the studied system. The information containing in different data sets is initially integrated by creating a combined matrix of aggregated profiles. The combined matrix is then passed to the k-means algorithm for subsequent analysis. Note that each one of the three partitioning algorithms can be applied in the considered context. k-means is used since it is efficient and easier to implement in comparison to the other two. However, it is more sensitive to outliers than k-medoids and k-medians. On the other hand, the latter two are more resource demanding.

B. Estimation of the Number of Clusters

The partitioning algorithms such as k-means and its variants discussed above, require k, the number of clusters to be known in advance. The main challenge in solving problems involving real-world data sets however, is to determine the optimal number of clusters. One solution to tackle this problem is to build a clustering model with a range of values for k and then evaluate the quality of the generated clustering partitions. For example, different internal cluster validation indices can be applied to recognize the optimal clustering solution. In this study, we take the advantages of two methods for identifying the optimal k: Elbow method and Silhouette Index.

Elbow method visualizes different numbers of k against a scoring parameter, for example, sum of squared distances, from each data point to its assigned centroid. Considering the complete plot, the point of inflection of the curve, *elbow*, can be recognized as the optimal k.

Silhouette Index (SI) [20] is a measure for computing the average closeness of each data point, to its neighbouring data points in the same cluster to the data points in the adjacent clusters. The data point is correctly clustered when its SI score is close to 1 and considered to be between two neighboring clusters in case of 0. Negative scores imply that the data point is misclassified and assigned to a wrong cluster.

In this study, k-means is conducted for each value of k in the interval [2, 10]. The corresponding scores of the Elbow index and SI generated by the different k are depicted as the function of k. The optimal k is the value at which a significant local change in the value of both methods observed.

C. Proposed Approach

In this study, we propose a multi-view data analysis approach that can be used for modelling and monitoring smart control valve behaviour. A particular control system is usually monitored in time via multiple sensors capturing measurements of different natures (e.g., operational parameters, contextual factors, and performance indicators). This will result in collecting measurements of several different characteristics that each represents part of the relevant information about the system behaviour and performance. Data analysis can benefit not only by considering data from all these observations but also by treating properly each different view about the studied system.

Let us assume that we have access to data of N characteristics of the control system under study. Consequently, the data corpus is composed of N different data sets one per monitored characteristic and each individual data set is composed of n daily time-series profiles.

The four main steps of the proposed multi-view data analysis approach are formally explained below. In addition, the approach is schematically illustrated in Figure 1.

1) *Multi-view interpretation of control valve system data*
   Analyse the available data attributes (N in total) and separate them into different groups with respect to the information they provide about the studied system, e.g., we can consider operational parameters, contextual factors and performance indicators.
2) *Model different control system operating modes*
   (a) Analyse the operational view data and select a subset of p_1 features that provides a comprehensive view of the system operational aspects.
   (b) Create integrated daily profiles by using the selected features, i.e., each integrated daily profile is a p_1-dimensional vector that consists of the aggregated values of the selected features.
   (c) Cluster the integrated daily profiles (n in total) into k disjoint groups.[1]
      - Each cluster is regarded as a different TOM of the control valve system, i.e., the system behaviour is modelled by k different TOMs.
      - Each TOM (cluster) is represented by a p_1-dimensional feature vector (behavioural signature), where the value of the i-th feature is set to be the average value over the corresponding values of all daily profiles (instances) assigned to the cluster.
3) *Annotate each TOM by using the remaining data views*
   (a) Analyse the contextual and performance data
      - Select a subset of p_2 features that describes the most relevant contextual factors affecting the system operational behaviour and performance.

[1]The integrated daily profiles can be partitioned by using k-means where the dissimilarity between any two daily profiles is computed by using Euclidean distance.

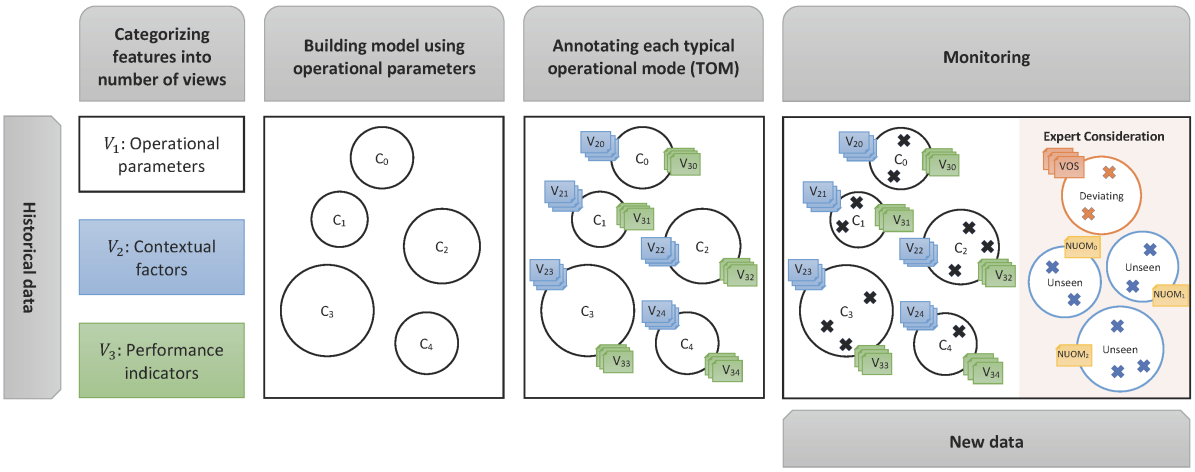

Fig. 1. Schematic illustration of the proposed approach. Observed data points during the monitoring step are indicated by "×", where their colour represent the nature of the recognized behaviour: *Typical* operational behaviour is in black, *Unseen* behaviour is in blue, and *Deviating* behaviour is the one in orange.

- Select a subset of p_3 indicators that provides the most realistic picture of the system performance.
- Each TOM is annotated by a p_2-dimensional contextual vector and a p_3-dimensional performance vector. Each label in any of these vectors is the average value over the corresponding labels of all daily profiles associated with this mode.

(b) The built multi-view model can be described as a $k \times (p_1 + p_2 + p_3)$ matrix (table), with the different behavioural modes corresponding to the rows and the different features (system characteristics) corresponding to the columns of the matrix and each cell (i, j) contains the representative value of feature j for mode i. Note that the built model explains system TOMs, their linked context, and expected performance.

4) *Context-aware monitoring of the system behaviour*

(a) The model built in Step 3 may be used for context-aware monitoring of the system behaviour by analysing the newly collected data and identifying deviating behaviour on daily basis.

(b) The built model may additionally be used for labelling a new daily profile with a specific mode. For example, three different scenarios (typical, deviating and, unseen) can be considered in this context:

- The feature (operational and contextual) values of the new daily profile are very close to some of the identified *typical* modes and the performance indicator values are as expected.
- The feature values of the new daily profile are very close to some of the identified modes, but the performance indicator values are very different from the expected ones, i.e., *deviating* behaviour.
- The feature values of the new daily profile are not close to any of the identified modes, i.e., a new *unseen* behavioral mode.

Data collected for a predetermined monitoring period (especially, those daily profiles annotated as representing deviating behaviour) can be discussed with the domain experts to further refine the labelling. The refined labeled data set can then be analysed and used to update the behavioural model, i.e., the proposed approach provides an opportunity for continuous improvement of the built model when new information becomes available.

IV. Experimental Design

A. Data

The data used in our experiments consists of several time-series data sets. Those have been collected from different types of sensors of an HVAC&R system. The data is unlabeled and covers one year time period, i.e., time-series data sets have been collected between 1^{st} of January 2019 and 1^{st} of January 2020. In addition, data is anonymized to protect and secure customers' privacy. Table I shows the list of all available features with their units in the data set. As a result of the discussions we have had with the domain experts, four new features are added that can provide additional information about the operational behaviour, performance, and contextual aspects of the control valve system. These new features are:

1) PD: A difference between primary supply temperature (PST) and primary return temperature (PRT) is referred to as primary delta (PD). The difference shows how much heat is used by a building, the larger the difference is better. The PD is computed as follows:

$$PD = PST - PRT \tag{1}$$

2) SE: A performance of a sub-station [2] can be measured in terms of efficiency using features (measurements) from

[2]A sub-station is an equipment that makes the water temperature and pressure at the primary side of the system suitable for the secondary side.

both the primary and the secondary sides. The sub-station efficiency (SE) is calculated as follows:

$$SE = \frac{PD}{PST - SRT}, \quad (2)$$

where PD is the temperature difference between supply and return water from the primary side (see eq. 1), PST is the primary supply temperature, and SRT is the secondary return temperature. The SE ranges between 0 to 1. A well-performed sub-station has a SE close to 1 (100%), however, due to the generation of domestic hot water, it can go higher than 1 [21].

3) RWB: The ratio of the number of weekends to the number of business days in a given cluster.
4) RCA: The ratio of the cardinality of a given cluster to the cardinality of the whole data set, i.e., the relative size of the cluster with respect to the other clusters.

TABLE I
FEATURES INCLUDED IN THE DATA SET

No.	Acronyms	Feature name	Units
1	PST	Primary Supply Temperature	°C
2	PRT	Primary Return Temperature	°C
3	**SST**	Secondary Supply Temperature	°C
4	**SRT**	Secondary Return Temperature	°C
5	**PHL**	Primary Heat Load	kW
6	**OT**	Outdoor Temperature	°C
7	**VOM**	Valve Openness Mean	%
8	**VOS**	Valve Openness Standard Deviation	%
9	PF	Primary Flow	m^3/h
10	PE	Primary Energy	MWh
11	PV	Primary Volume	m^3
12	PD	Primary Delta	°C
13	**SE**	Sub-station Efficiency	%
14	**RWB**	Ratio of weekends to business days	%
15	**RCA**	Ratio of number of days in a cluster to the total number of days in data	%

Features 1 to 11 (above horizontal line) present the measurements collected from the sensors. The remaining four features are created and added to the study. The selected features are shown in bold.

Since the different features are collected within different time intervals, after having discussions with the domain experts and in order to be consistent, the data for each feature is aggregated into one hour resolution. Note that the daily profiles (average daily values of the selected features) are used by the proposed approach to model the different behavioural modes of the studied control valve system. Hourly profiles of the same features are used to create zoomed views, which can supply the domain expert with further details and opportunities for comparing and understanding different behavioural modes of the system.

B. Data Preprocessing

Real-world data sets often contain missing values that requires careful attention to be treated. Mean substitution, hot-deck imputation [22], regression analysis, and multiple imputation [23] are some of the examples of imputation techniques. The suitable imputation method should always be selected based on the nature of data. The data used in our experiments have missing values only in the last four days of 2019 due to the fact that in these dates the system was shutdown. Therefore, it was not necessary to apply any sophisticated imputation technique on our data sets. Those four days were only removed from the data sets.

In this study, to identify and remove the outliers in each feature, Hampel filter [24], a method based on median absolute deviation (MAD) estimation is applied. The filter belongs to the class of three-sigma rules of statistics which makes it robust against outliers. In addition, data is scaled by applying z-score normalization on each feature (data set) to have a mean equal to zero and standard deviation equal to one.

We assess the control valve system behaviour during the heating season. Therefore, those weeks with average outdoor temperatures below a certain threshold are selected for our analysis. The threshold for this selection is set to be 10 °C due to the discussion we have had with the domain experts.

We initially analyse the available data features and partition them in three distinctive views as follows:

1) The features that represent the operational behaviour of the control valve system are: PST, PRT, SRT, SST, PHL, PF, PE, PV, and PD. Among these features, SST, SRT, and PHL are selected to model the system TOMs. SST and SRT are used since they provide information about the secondary side behaviour and how efficiently the provided heat is consumed. PHL represents the primary side behaviour, which has a high correlation with PV, PF, PE, and OT.
2) The control valve system performance can be evaluated by these three indicators: VOM, VOS, and SE.
3) The contextual factors are represented by the features: OT, RWB, and RCA.

C. Experimental Scenario

The monitoring (streaming) scenario is modelled by dividing the data set into two parts, training and monitoring. The training data consists of 70% of the days in each month that belongs to the heating season. The remaining 30% of the data represents the new arrival data which is used for monitoring. In addition, a domain expert has provided us with information about 49 dates (12^{th} of March to 29^{th} of April 2019) with abnormal system behaviour. The corresponding dates are placed in the second data set. In order to reduce any bias introducing by this process, five data set couples (where each data set couple contains 112 training and 80 monitoring daily profiles, respectively) by randomly separating the daily profiles in two sets are created. Note that the training data sets only contain unlabeled data which is used to model the system behaviour.

V. RESULTS AND DISCUSSION

In this section, we present and discuss the results obtained by applying the proposed approach on one of the five data

TABLE II
THE BUILT MULTI-VIEW MODEL TOGETHER WITH FIVE TOMS, THEIR LINKED CONTEXT AND EXPECTED PERFORMANCE

TOM	SST	SRT	PHL	OT	RWB	RCA	VOM	VOS	SE
0	42.64	36.81	23.55	5.67	30	30	14.16	±0.51	95
1	46.88	37.92	39.76	1.51	40	18	17.94	±0.49	96
2	51.94	40.18	51.36	-2.96	18	15	18.96	±0.64	97
3	36.21	32.86	13.68	10.25	21	12	11.97	±0.50	89
4	46.00	39.84	23.41	2.64	32	25	13.67	±0.50	96

Note. The features show average values for each TOM. The unit for SST, SRT, PHL, and OT is °C. VOM, VOS, SE, RWB, and RCA are expressed in %. For the full form of each feature see Table I.

set couples (see Section IV-C). The results generated on the remaining data set couples are similar to the presented ones.

As it is explained in Section III-C, the available features are initially partitioned into three different views. In addition, three features are selected as representatives for each view (see Section IV-B). The next step is to identify the system TOMs by applying k-means to analyse the integrated daily profiles of the three operational parameters. In the studied experimental scenario, five distinctive TOMs are recognized (see Section III-B). Each TOM is additionally annotated by its context and performance through analysing the respective data views (see Section III-C for details). Table II shows the built multi-view control valve system model that represents the TOMs of the system, their linked contextual factors, and expected performance.

In order to supply the domain experts with additional support in the analysis and understanding of the different TOMs, zoomed views of the studied features can be created. Note that the zoomed views meant to provide an overall status of each TOM in a 24-hour period. Figure 2 shows the zoomed views of four features: SE, OT, VOM, and PHL. For example, if we compare TOM 1 and TOM 2 for the two time slots, 00:00am-02:00am and 10:00pm-11:00pm, we can notice that the measured values both for SE and VOM are decreasing. This may be due to heat load (PHL) reduction, specifically domestic hot water consumption, in those time periods.

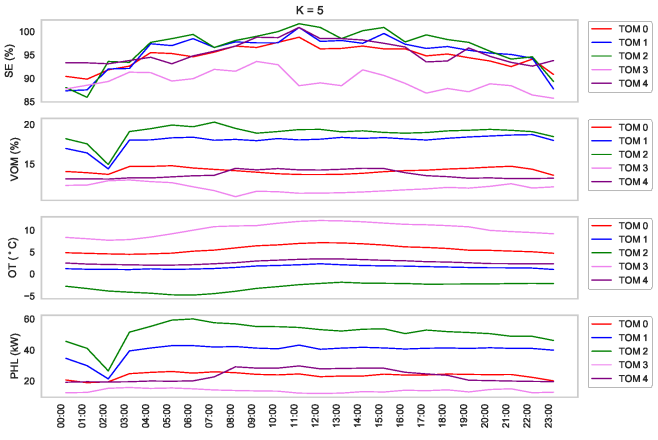

Fig. 2. Time-series plots of the operating, contextual and performance characteristics' zoomed views

The comparison of TOMs shows that when OT decreases, VOM including VOS, SE, and PHL increase. For example, the lowest OT is linked to TOM 2 and respectively VOM, VOS, SE, and PHL have the highest values for this TOM. On the other hand, by comparing TOMs 0 and 4, we observe that although TOM 4 has a lower OT than TOM 0, its VOM is slightly lower than the detected in TOM 0. One reason for such behaviour can be that the average standard deviation of OT in TOM 0 is higher than TOM 4 as it is shown in Table III.

TABLE III
COMPARISON OF THE TOMS USING OT STANDARD DEVIATION

TOM	OT standard deviation (°C)
0	±1.34
1	±1.05
2	±2.07
3	±1.81
4	±1.04

More detailed information regarding the comparison of TOMs 0 and 4 is provided in Table IV. As one can notice the number of daily profiles associated with TOM 0 is slightly higher than of those belonging to TOM 4. In addition, TOM 0 contains daily profiles from six different months while in TOM 4, only daily profiles from three of those months (October, November, and December) are presented.

TABLE IV
CONTEXTUAL STATISTICS ABOUT TOM 0 AND TOM 4

TOM	Month	Number of days	OT (°C)
0	February	10	1.44
	March	3	1.83
	May	2	1.98
	October	5	1.64
	November	8	1.01
	December	5	0.83
	Total	33	
4	October	2	1.14
	November	12	1.11
	December	14	0.97
	Total	28	

TABLE V
RESULTS FROM MONITORING THE DAILY PROFILES USING THE PROPOSED METHOD

No.	Date	Day of Week	OT (°C)	Outlier type	Reason	Action
1	2019-03-12	Tue	-1.55	Deviating	VOS	Analyse and Diagnose
2	2019-03-13	Wed	2.19	Unseen	$NUOM_0$	Update Model, Expert Consideration
3	2019-03-14	Thu	4.70	Unseen	$AUOM_0$	Expert Consideration
4	2019-03-15	Fri	3.50	Unseen	$AUOM_0$	Expert Consideration
5	2019-03-16	Sat	2.81	Unseen	$AUOM_0$	Expert Consideration
6	2019-03-17	Sun	3.91	Unseen	$AUOM_0$	Expert Consideration
7	2019-03-18	Mon	4.03	Unseen	$AUOM_0$	Expert Consideration
8	2019-03-19	Tue	2.59	Unseen	$AUOM_0$	Expert Consideration
9	2019-03-20	Wed	4.43	Unseen	$AUOM_0$	Expert Consideration
10	2019-03-21	Thu	7.13	Unseen	$AUOM_0$	Expert Consideration
11	2019-03-22	Fri	6.63	Unseen	$AUOM_0$	Expert Consideration
12	2019-03-23	Sat	7.77	Deviating	VOS	Analyse and Diagnose
13	2019-03-24	Sun	7.23	Deviating	VOS	Analyse and Diagnose
14	2019-03-25	Mon	5.38	Unseen	$AUOM_0$	Expert Consideration
15	2019-03-26	Tue	3.98	Unseen	$AUOM_0$	Expert Consideration
16	2019-03-27	Wed	6.04	Unseen	$AUOM_0$	Expert Consideration
17	2019-03-31	Sun	5.72	Unseen	$AUOM_0$	Expert Consideration
18	2019-04-01	Mon	6.13	Unseen	$AUOM_0$	Expert Consideration
19	2019-04-02	Tue	6.04	Unseen	$AUOM_0$	Expert Consideration
20	2019-04-03	Wed	7.01	Unseen	$AUOM_0$	Expert Consideration
21	2019-04-04	Thu	8.15	Deviating	VOS	Analyse and Diagnose
22	2019-04-05	Fri	6.02	Unseen	$AUOM_0$	Expert Consideration
23	2019-04-06	Sat	8.22	Deviating	VOS	Analyse and Diagnose
24	2019-04-07	Sun	7.29	Deviating	VOS	Analyse and Diagnose
25	2019-04-08	Mon	2.91	Unseen	$AUOM_0$	Expert Consideration
26	2019-04-09	Tue	1.91	Unseen	$AUOM_0$	Expert Consideration
27	2019-04-10	Wed	2.48	Unseen	$AUOM_0$	Expert Consideration
28	2019-04-11	Thu	3.13	Unseen	$AUOM_0$	Expert Consideration
29	2019-04-12	Fri	3.35	Unseen	$AUOM_0$	Expert Consideration
30	2019-04-13	Sat	4.79	Unseen	$AUOM_0$	Expert Consideration
31	2019-04-14	Sun	4.85	Unseen	$AUOM_0$	Expert Consideration

Note. The OT values show the average outdoor temperature of each day. VOS: valve openness standard deviation, NUOM: newly observed unseen operating mode, and AUOM: already observed unseen operating mode

Although, the detection of deviating behaviour is crucial for system maintenance, providing additional information about eventual reason(s) for such behaviours are also as much essential. Due to the complexity of the control valve system and the availability of a high number of features, finding the cause of deviating behaviour is extremely difficult and time consuming for the domain experts. Therefore, we believe that the explainable results provided by our multi-view model can work as recommendations for the domain experts to facilitate them in the analysis and interpretation of the system's behaviour and performance. The additional information may supply domain experts with the opportunity for better understanding and gaining additional knowledge about the system operating modes. Another advantage of the proposed method is that it uses historical data for the initial modelling of different TOMs of the control valve system. However, the built model is not static, since the monitoring step of our approach can be considered as a continuous learning process that provides the opportunity for non-stopping refinement and updating of the model as more data becomes available.

As mentioned earlier the new arrival data set contains 80 daily profiles. Applying the proposed approach to this data set leads to identifying 32 daily profiles as typical behaviour. More specifically 5 profiles are classified as TOM 0, 11 as TOM 1, 3 as TOM 2, 6 as TOM 3, and 7 as TOM 4. The remaining daily profiles are classified as follows: 23 identified as *Deviating* while 25 recognized as *Unseen*. In the case of *Deviating* 3 profiles are labeled based on SE indicator and the rest are annotated based on VOS.

Table V lists part of the obtained results during the monitoring step. In this table, two types of outliers can be seen which are respectively labeled as *Deviating* and *Unseen*. The daily profiles labeled with *Deviating* are those that have very similar values, based on domain-specific thresholds[3], for the operational parameters and contextual factors to one of the TOMs. However, their performance indicators' values in comparison to a domain-specific threshold (which is defined similarly to the other set thresholds) are different. For instance,

[3]These thresholds are set to be less than or equal to 2 times the standard deviation of each operational parameter and contextual factor, respectively.

in Table V row numbers 1, 12, 13, 21, 23, and 24 are detected as *Deviating* due to having different VOS. The daily profiles annotated with *Unseen*, on the other hand, are not similar to any of the TOM profiles given in Table II. This category can further be divided into two groups. Namely, newly observed unseen operating mode (NUOM) and already observed unseen operating mode (AUOM). Note that the daily profiles belong to *Unseen* category can be clustered based on their similarities into a number of clusters. All NUOMs observed for a predefined monitoring period need to be analysed and used for updating the current TOMs.

The obtained results show that the proposed approach has managed to detect the days with abnormal behaviours supplied by the domain experts. During that time period, due to a malfunction on the primary side, the system was not working properly, which had a high impact on the operational behaviour of the control valve system. According to the discussed scenarios in Step 4 of the proposed approach (see Section III-C) some of those days are annotated as *Deviating* and the remaining days as *Unseen*. Note that any issues in either sides of the main system (primary or secondary) can affect the sub-systems of an HVAC&R including the control valve system. These systems are designed to meet consumers' demands despite possible faults degrading their performance. However, in a long term, this can lead to higher costs in terms of energy consumption and maintenance. Therefore, FDD systems capable of early detection and classification of faults, similar to the approach proposed in this study, are highly required for the discussed domain. The executable algorithm and the obtained experimental results are available at GitHub[4]. Data is not public since it is an asset of the company.

VI. Conclusion and Future Work

In this study, we have proposed a multi-view data analysis approach for modelling and monitoring smart control valve system behaviour. The proposed approach has been evaluated on real-world data sets. The obtained results have demonstrated the robustness of the proposed approach in analysing and identifying the control valve system deviating behaviour.

For future work, we are interested in studying the transferability of the developed model for behavioural monitoring problems of other smart systems, e.g., heat exchanger efficiency in ventilation systems. In addition, two future directions can be considered in the context of this study:

1) The monitoring step of the proposed method can be considered as a continuous learning process. Namely, the monitored daily profiles can be annotated by comparing them with the known typical behavioural modes. This information can be analysed in predefined time intervals and used for continuous updating and improvement of the built model. The implementation of the updating step of the approach is part of our future plans.
2) The proposed method is capable of performing fault detection by identifying the deviating behaviour of the control valve system. The latter can happen due to different faults or issues such as valve degradation, unsuitable size of the valve, etc. The diagnostic functionality is not implemented in the current version of the approach and is also part of our future plans.

[4]https://github.com/amirme/ICMLA20

References

[1] "SmartThings," www.smartthings.com, [Online; accessed 26-July-2020].
[2] "Vera," www.getvera.com, [Online; accessed 26-July-2020].
[3] "openHAB," www.openhab.org, [Online; accessed 26-July-2020].
[4] S. Katipamula and M. R. Brambley, "Methods for fault detection, diagnostics, and prognostics for building systems—a review, part 1," *Hvac&R Research*, vol. 11, no. 1, pp. 3–25, 2005.
[5] ——, "Methods for fault detection, diagnostics, and prognostics for building systems—a review, part 2," *Hvac&R Research*, vol. 11, no. 2, pp. 169–187, 2005.
[6] D. Anderson, L. Graves, W. Reinert, J. Kreider, J. Dow, and H. Wubbena, "A quasi-real-time expert system for commercial building hvac diagnostics," *ASHRAE Transactions*, vol. 95, no. CONF-890609–, 1989.
[7] P. Usoro, I. Schick, and S. Negahdaripour, "An innovation-based methodology for hvac system fault detection," 1985.
[8] M. McKellar, "Failure diagnosis for a household refrigerator," *Master's thesis, School of Mechanical Engineering, Purdue University, West Lafayette, Indiana*, 1987.
[9] L. Stallard, "Model based expert system for failure detection and identification of household refrigerators," *Master's thesis, School of Mechanical Engineering, Purdue University, Indiana*, 1989.
[10] W. Kim and S. Katipamula, "A review of fault detection and diagnostics methods for building systems," *Science and Technology for the Built Environment*, vol. 24, no. 1, pp. 3–21, 2018.
[11] N. Ren *et al.*, "Fault diagnosis strategy for incompletely described samples and its application to refrigeration system," *Mechanical Systems and Signal Processing*, vol. 22, no. 2, pp. 436–450, 2008.
[12] M. Najafi, D. M. Auslander, P. Haves, and M. D. Sohn, "A statistical pattern analysis framework for rooftop unit diagnostics," *HVAC&R Research*, vol. 18, no. 3, pp. 406–416, 2012.
[13] J. Cui and S. Wang, "A model-based online fault detection and diagnosis strategy for centrifugal chiller systems," *Int'l. J. of Thermal Sciences*, vol. 44, no. 10, pp. 986–999, 2005.
[14] J. C.-M. Yiu and S. Wang, "Multiple armax modeling scheme for forecasting air conditioning system performance," *Energy Conversion and Management*, vol. 48, no. 8, pp. 2276–2285, 2007.
[15] P. R. Armstrong, C. R. Laughman, S. B. Leeb, and L. K. Norford, "Detection of rooftop cooling unit faults based on electrical measurements," *HVAC&R Research*, vol. 12, no. 1, pp. 151–175, 2006.
[16] Z. Du, X. Jin, and L. Wu, "Fault detection and diagnosis based on improved pca with jaa method in vav systems," *Building and Environment*, vol. 42, no. 9, pp. 3221–3232, 2007.
[17] M. Kim *et al.*, "Cooling mode fault detection and diagnosis method for a residential heat pump," *NIST Special Publication*, vol. 1087, 2008.
[18] B. Fan, Z. Du, X. Jin, X. Yang, and Y. Guo, "A hybrid fdd strategy for local system of ahu based on artificial neural network and wavelet analysis," *Building & Environment*, vol. 45, no. 12, pp. 2698–2708, 2010.
[19] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. of 5th Berkeley symposium on math. statistics and probability*, vol. 1, no. 14, 1967, pp. 281–297.
[20] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
[21] S. Abghari, *et al.*, "Higher order mining for monitoring district heating substations," in *2019 IEEE Int. Conf. on Data Science and Advanced Analytics*, pp. 382–391.
[22] B. L. Ford, "An overview of hot-deck procedures," *Incomplete data in sample surveys*, vol. 2, no. Part IV, pp. 185–207, 1983.
[23] D. B. Rubin, "Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse," in *Proc. of the survey research methods section of the American Stat. Assoc.*, vol. 1, 1978, pp. 20–34.
[24] F. R. Hampel, "A general qualitative definition of robustness," *The Annals of Mathematical Statistics*, pp. 1887–1896, 1971.