DISSERTATION

Fault Detection and Diagnosis in Building Energy Systems

Submitted at the Faculty of Electrical Engineering and Information Technology, Technische Universität Wien in partial fulfillment of the requirements for the degree of Doktor der technischen Wissenschaften (equals Ph.D.)

under supervision of

Em. O. Univ. Prof. Dipl.-Ing. Dr. techn. Dietmar Dietrich
Institute number: E384
Institute of Computer Technology
Technische Universität Wien

and

Priv.-Doz. Dr. Nysret Musliu
Institute number: E184
Institute of Information Systems
Technische Universität Wien

Co-supervised by

Dipl.-Ing. Dr. Gerhard Zucker
Energy Department
Austrian Institute of Technology

by

Usman Habib
Matr.Nr. 1325027
Längenfeldgasse 22/16, 1120, Vienna, Austria

April, 2016.

Kurzfassung

Das massive Wachstum an Energiebedarf erfordert Gegenmaßnahmen in Form von Energieeinsparungen, wobei Gebäude hier eine wesentliche Rolle spielen, da sie einen signifikanten Anteil am Energieverbrauch haben. Öffentliche, Wohn- und kommerzielle Gebäude tragen 40% in Europa und 41% in den USA zum Gesamtenergieverbrauch bei. Studien zeigen, dass davon ein Großteil in Heizung, Lüftung und Klima (HLK) geht. Um diesen Energieverbrauch zu reduzieren ist es erforderlich, dass der Betrieb der HLK-Komponenten nicht nur effizient, sondern auch möglichst fehlerfrei erfolgt. Das Verhalten der unterschiedlichen Komponenten wird sensorisch erfasst, um dadurch Fehler erkennen und diagnostizieren zu können. Die anfallende Datenmenge macht eine durchgehende Analyse mittels einfacher Visualisierungen in Hinblick auf Datenfehler sehr unpraktikabel. Dafür sind Methoden zur automatisierten Fehlererkennung erforderlich.

Diese Arbeit konzentriert sich auf automatische Erkennung und Diagnose von Fehlern in Gebäudeenergiesystemen unter Verwendung von Datenanalyse- und Maschinenlernalgorithmen mit dem Ziel, den Konfigurationsaufwand sowie das erforderliche Domänenwissen möglichst gering zu halten. Im ersten Schritt wurde ein Workflow zur Datenbereinigung und Betriebsdatenanalyse erarbeitet, um so die Zuverlässigkeit der Betriebsdaten zu erhöhen. Durch diesen Workflow wurde eine automatisierte Betriebszyklus-Erkennung ebenso wie eine Datenvalidierung über Grundprinzipien wie Massen- oder Energieerhaltung erreicht und außerdem einige Visualisierungen für die erste Analyse sowie die Erkennung von Daten-Ausreißern und Datenlöchern erarbeitet, um die Datenqualität zu erhöhen. Zusätzlich wurden Methoden des Mustervergleichs angewendet, um verschiedene Datenanomalien während der Datenaufzeichnung zu detektieren.

Um unterschiedliche Betriebsmuster in den Betriebszyklen von Energiesystemen identifizieren zu können, wurde Daten-Clustering verwendet, das die Betriebsdaten als diskrete Muster repräsentiert. Diese Muster wurden aus den Betriebszyklusdaten extrahiert und dem Clustering-Algorithmus als Eingänge gegeben, der diese nach den Betriebsmustern gruppiert. Mithilfe statistischer Analyse wurde die optimale Cluster-Anzahl ermittelt, indem der geringste Unterschied zwischen Intra-Cluster Elementen und größter Abstand zwischen Inter-Cluster-Elementen ermittelt wurde.

Der in dieser Arbeit erstellte Workflow wurde auf Betriebsdaten unterschiedlicher Energiesysteme angewendet. Die vorgeschlagene Methodik hat sich als sehr hilfreich bei der Extraktion von Information für die Datenanalyse erwiesen und hat Fehler erfolgreich diagnostiziert. Der Workflow hat somit zur Domäne der Fehlererkennung und -diagnose im Bereich von Gebäudeenergiesystemen beigetragen.

Abstract

The rapid growth in the global energy demand necessitates its efficient use in order to reduce energy consumption. Buildings share the major portion of the total energy consumption of the world: the energy demands of residential, public and commercial buildings together constitute a 40% of the European and 41% of the US total energy consumption. Investigations show that the main component of energy consumption in a typical building is the Heating, Ventilation and Air-conditioning (HVAC) facility. To reduce energy consumption, it is thus very important that the operation of HVAC components is not only efficient but also without any fault. In order to observe the various types of behavior of energy systems in buildings, these components are monitored using sensors that can be used to detect and diagnose the different types of faults in the HVAC components. The size of the data makes it difficult and very inefficient to continuously analyze the data with simple visualizations for detection and diagnosis of faults in the data. Therefore, it is required to develop tools and techniques for automatic detection and diagnosis of faults in the data.

This research focuses on automatic detection and diagnosis of faults in the energy systems of buildings using data analytics and machine learning algorithms using a minimum of required configuration setup and domain knowledge. In the first step a work-flow for data sanitation and analysis of operation data has been proposed for increasing reliability in the operation data of building related energy systems. The main tasks achieved in the sanitation work-flow are automatic detection of duty cycles (operational), validation of data using first principles, various visualizations for basic analysis, outliers detection in the data, and missing data imputation, to ensure the data quality for the analysis of the energy system. Moreover, a solution has been proposed using pattern matching method, to detect various anomalies in the data that have been added while monitoring the energy systems.

Furthermore, use of clustering is suggested for finding the various patterns of the operational cycles in data of the energy systems in building. The data of the operational cycles is represented in a discretized pattern form, extracted from the operational cycles data, which has been provided as an input to the clustering algorithm. The clustering algorithm groups the duty cycles according to their operation pattern. The optimal number of clusters was chosen by using the gaps statistical analysis that finds the minimum difference between the intra-cluster elements and maximum difference between the inter-cluster elements.

The proposed work-flow has been applied using operational data of different energy systems. The suggested methodology has been helpful in extracting useful information for analysis and successfully detected and diagnosed faults. The proposed work-flow leads to a novel contribution in the domain of fault detection and diagnosis in the energy systems of building.

Acknowledgements

This research work was carried out at Austrian Institute of Technology (AIT), Energy Department as a part of project "extrACT", project number 838688, which was partly funded by the Austrian Funding Agency in the funding programme e!MISSION.

I would like to thank my supervisor Professor Dietmar Dietrich, who was very cooperative, and without his guidance my research work will not have finished in time. I would like to extend my gratitude to Professor Nysret Musliu as his input always helped a lot to improve my research work and has been instrumental in my successful completion. I am also thankful to my co-supervisor Gerhard Zucker, who always helped and supported me in every stage of the research. His motivation, encouragement, patience and belief in my abilities made this research work a success. My special thanks to Professor Khizar Hayat whose critical reviews and suggestions helped me a lot to polish my work.

I would also like to mention Austrian Institute of technology (AIT), where I got a place for study and was fully facilitated by the staff. The support of the people at AIT, especially Max Blöchle, Jasmine Malinao, Florian Judex and Tim Selke has always helped me to improve my skills. I am feeling lucky to be a part of such prestigious research institute for three years and have taken full advantage of the experts dealing practical problems of the world. The discussions with these experts were the most valuable thing I have gained. The experience and training of AIT will always help me in future projects and will be positive addition to my experience.

This research would not be possible without the funding provided by Higher Education Commission (HEC) of Pakistan and partner agency OeAD Austria. I had a very good experience in dealing with the OeAD office and have found them very cooperative.

I am thankful to Almighty Allah for everything. At the end, I would like to pay my special gratitude to my parents and my aunt Abida for their prayers and support. Not to forget, my better half Mrs. Saima khan whose support in difficult times has always kept me calm and relieved from stress, and my son Muhammad Saad Usman, whose smile and naughty looks always gave me the strength to face every difficulty of life.

Table of Contents

Abbreviations

ANN	Artificial Neural Network
ARIMA	Auto Regressive Integrated Moving Average
BoWR	Bag of Word Representation
CoP	Coefficient Of Performance
CAV	Constant air volume
DEC	Desiccant Evaporative Cooling
DTW	Dynamic Time Warping
$E6$	High Temperature (HT) electricity consumption meter
$E7$	Medium Temperature (MT) electricity consumption meter
$E8$	Low Temperature (LT) electricity consumption meter
EM	Expectation Maximization
FAR	False Alarm Rate
FDD	Fault Detection and Diagnosis
HT	High Temperature
HVAC	Heating Ventilation and Air Conditioning
IEA	International Energy Agency
KNN	K-Nearest Neighbors
LT	Low Temperature
MT	Medium Temperature
PoD	Probability of Detection
$PR6$	Pressure in High Temperature cycle
$PR7$	Pressure in Medium Temperature cycle
$PR8$	Pressure in High Temperature cycle
$Q12\_KW$	Medium Temperature cycle Energy consumption reading
$Q12\_m3h$	Medium Temperature cycle Flow (water) reading
$Q6a\_KW$	High Temperature cycle Energy consumption reading
$Q6a\_m3h$	High Temperature cycle Flow (water) reading
$Q7\_KW$	Low Temperature cycle Energy consumption reading
$Q7\_m3h$	Low Temperature cycle Flow (water) reading
SAX	Symbolic Aggregate Approximation
SHC	Solar Heating and Cooling
SHDC	Solar Heat Driven Cooling
SVM	Support Vector Machine
T_HTre	Temperature at return side of high temperature cycle
T_HTsu	Temperature at supply side of high temperature cycle
T_LTre	Temperature at return side of low temperature cycle
T_LTsu	Temperature at supply side of low temperature cycle
T_MTre	Temperature at return side of medium temperature cycle
T_MTsu	Temperature at supply side of medium temperature cycle
VAV	Variable air volume

1 Introduction

Shelter is the basic necessity of humans, be it for housing, industry or social sector; health, education, entertainment and other government services. With high population densities in cities, buildings are required to be huge and multi-storey. In addition, there is always an upward trend for having a closely controlled working environment in different types of buildings such as hospitals, laboratories, and industrial facilities. However, the human comfort comes first and it is essential to provide a comfortable environment in buildings, in order to facilitate the people to live a healthy life. All this comes at the expense of the energy; a demand that is continuously escalating with the ever-evolving human comfort needs. All this boils down to the efficient use of energy. Therefore the energy systems researchers and experts are challenged to come up with new methods in order to make the buildings' operations energy efficient.

With buildings being one of the main culprits in high energy consumption, it is required to study their energy systems in detail. The energy system of a typical building consists of several subsystems such as heating, ventilation, air-conditioning and lighting. The analysis of a building's energy system is a complex task since the underlying subsystems are usually interdependent. For a detailed analysis, the behavior of the energy system is observed and recorded using sensors. One of the simplest methods is the use of data visualization tools followed by manual analysis in order to monitor the performance of buildings. This can be time intensive and there is a high chance of missing some important and interesting patterns in the data.

Nowadays, a lot of data is recorded during the monitoring of a building's energy systems. The benefit of saving a huge amount of data is lost due to the fact that it is humanly impossible to manually analyze such a huge amount of data in detail. As a result, most of the data is not evaluated at all despite having the potential of ascertaining the efficient operation of the system via rapid detection of errors and faults. At present, usually the detection of faults in data is restricted to the launch of alarms in situations of severe faults. In order to ensure the permanent monitoring of the data, it is required to have an implementation for the automatic interpretation of the data. This will help in the proper functioning of the energy system, as well as in the detection of faults.

Machine learning methods, like clustering, can be used to discover hidden patterns in the data that can be useful in fault detection and diagnosis and may further help in reducing the energy cost. Figure 1.1 elaborates the general process of how the different behavior patterns in the buildings energy system can be extracted using machine learning algorithms. This process can help to automatically evaluate the huge amount of the data stored as well as in detecting and diagnosing of faults as a consequence.

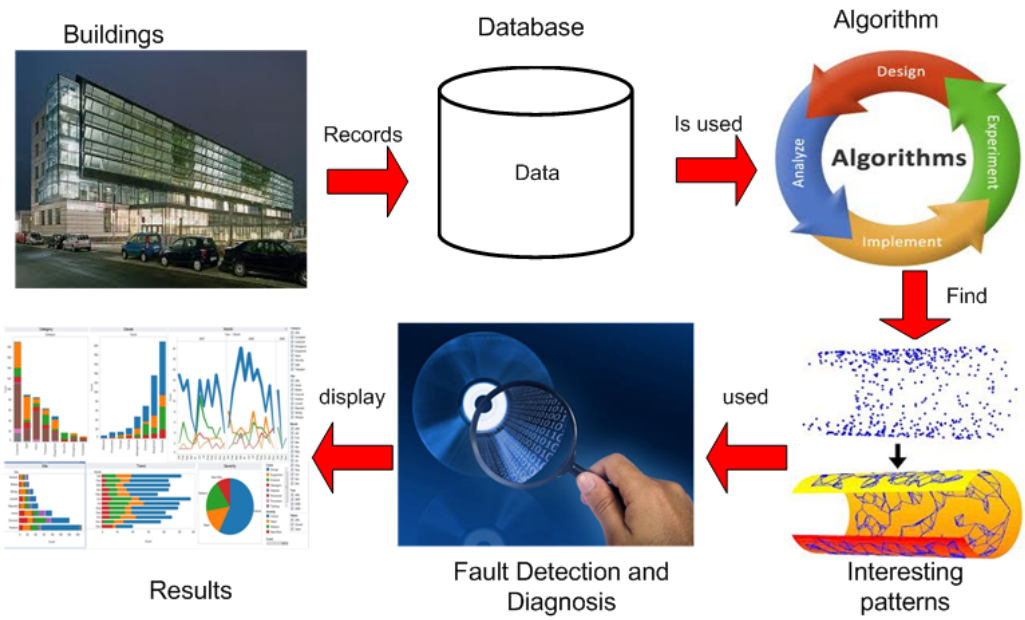


Figure 1.1: Process of Fault Detection and Diagnosis

1.1 Motivation

The energy consumption of the world and the projection shows upswing by nearly 50% from 2009 till 2035. The growth is mostly taking place in the developing economies outside the Organization for Economic Cooperation and Development (OECD), particularly in non-OECD Asia. The projected period shows the upward trend of 84% in total non-OECD energy use, as compared to a 14% increase in the developed OECD nations. It can be noticed in Figure 1.2 that China and USA are the prominent countries showing 20% and 19% of the total world energy consumption share of in 2010 [1], respectively.

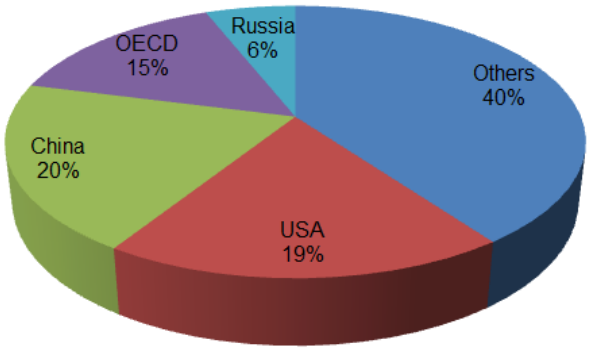


Figure 1.2: World energy consumption

Increasing energy prices and energy demand pose new challenges for modern society. Furthermore, the total energy consumption of the world can be divided into four main sectors i.e. transportation, industry, residential and commercial buildings . According to the Figure 1.3, buildings are one of the fastest growing energy consuming sectors. The buildings share of the energy consumption can be further categorized as commercial and residential buildings. In the year 2008, the

energy consumption share of the buildings sector was about 21% of the total world energy consumption as can be seen in Figure 1.3. Moreover, the energy consumption in the residential sector was about 14%, whereas, the non-residential sector accounted for 7% of total energy consumption of the world [1].

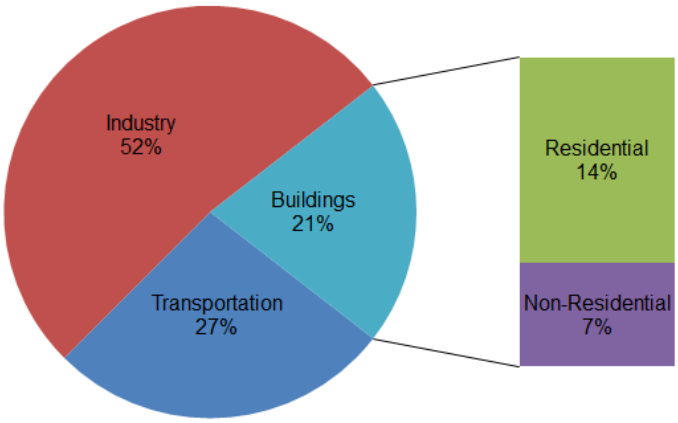


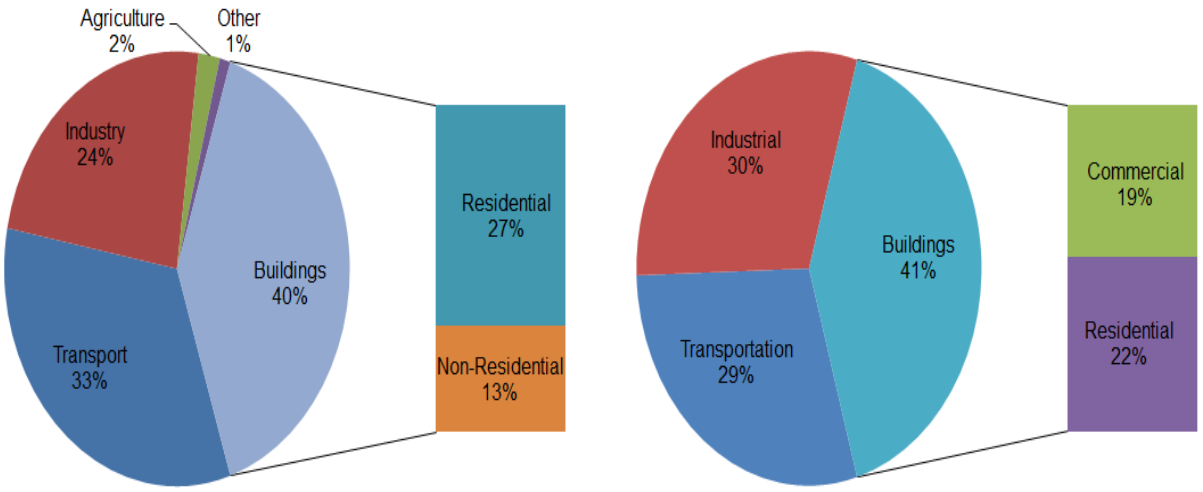
Figure 1.3: World energy consumption sector wise in 2008

The total energy consumed in residential and non-residential buildings was around 40% of the total Europe energy requirements in 2009 [2]. It is noteworthy to observe that the buildings sector is representing the key sector in European energy consumption with 40%, followed by transport with 33% and industry sector with 24%, as can be seen in Figure 1.4(a). The energy consumption trends from 1990 to 2009 display two main points i.e. increase by 50% in electricity and gas usage, whereas reduction in the usage of oil and solid fuels with 27% and 75% respectively. It is a matter of greater concern to see the statistics of over the last 20 years for energy use in buildings moving in upward direction with a rise from 400 Mtoe (Million tons of oil equivalent) to 450 Mtoe.

In 2009, European households share the major portion of energy consumption with 68% of the building sector of Europe. The increase in use of household's appliances is also evident with the 38% growth in the electricity consumption in the period of 20 years from 1990 to 2009. Moreover, the electricity consumption in non-residential buildings has an increase of about 74% during the same period of time. The statistics show that 50% of the energy is consumed by Offices, wholesale and retail trade buildings, whereas, the education and sports facilities consumed 18% of the energy.

In 2010, only USA building sector consumed 7% of the global energy consumption that is about 41% of the total energy consumed in USA as can be seen in Figure 1.4(b) [3]. The primary energy consumption sector in USA buildings are the residential sector covering 54% share of the building sector consumption and 22% of the total energy consumption of USA. Furthermore, the commercial buildings are representing the energy consumption of 46% in the USA total buildingsenergy consumption.As can be seen in Figure 1.4(b) the commercial buildings are consuming 18.9% of the total USA energy consumption.

One of the main reasons for the rise in the energy consumption of the buildings is because of the increase in the energy demands due to space heating, cooling, ventilation, and refrigeration requirements [4]. The energy consumption in buildings is directly connected to the energy demands of heating, ventilation and air-conditioning (HVAC) systems. The statistical data shows that HVAC is the biggest energy share in both the residential and non-residential sector energy



(a) Europe energy consumption sector wise

(b) US energy consumption sector wise

Figure 1.4: Europe and US energy consumption

consumption. The air-conditioning is mostly contributing to the building energy consumption and can range from 10% to 60%, depending on the building's type, as shown in different studies [5]. The statistics have shown that HVAC systems are significantly energy consuming devices, ranging from about 10% to 20% of total energy consumption in the developed countries [6].

Moreover, the data shows that the combined space heating and cooling of the buildings in the USA share the 50% of the total energy consumption as can be seen in Figure 1.5 [3]. The HVAC in residential sector consume 61% of the total residential energy consumption whereas, share41% of the total HVAC elements consumption. It is important to note that the percentage of new single family homes equipped with air conditioning has increased from 62% to 88% in the time period from 1980 to 2010. Similarly, equipment such as heat pump heating systems have also expanded its market share from 23% to 38% in the time span from 2001 to 2010.

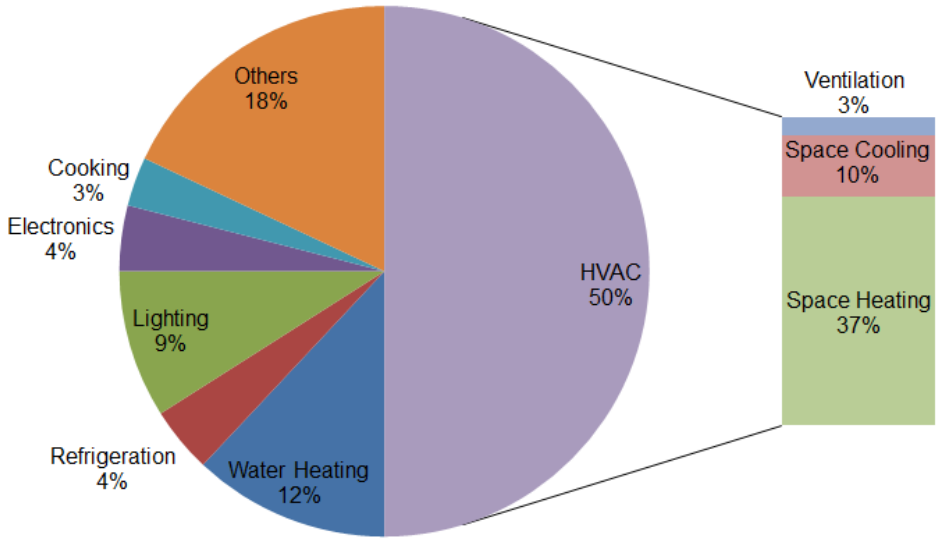


Figure 1.5: Building energy consumption sections

There are different factors that will cause increase the energy consumption in buildings such as growth in the population, global climate change, comfort requirement,and the number of hours spent inside buildings (about 90% of our whole life). The growth at an average rate of 1.1% per year from 2008 to 2035 is predicted in residential energy use. In the same way, the commercial sector growth is estimated to increase at an average rate of 1.5% per year from time span between 2008 and 2035 [4].

The failure of any energy component can lead to increased energy consumption and contribute to the discomfort of the people in the building. Furthermore, various results demonstrate that most of the faults can be detected and diagnosed earlier and can help in saving the energy e.g. in [7], the authors have shown the faults in simple rooftop air conditioner can be detected earlier then the decrease in capacity of 5% is attained. Therefore it is required to develop such methods that can automatically detect and diagnose faults, in order to save energy and cost of equipment as well.

1.2 Problem Statement

The processes involved in the operation of energy systems of the building are quite complex, and analyzing the behavior of such systems is very challenging. In order to capture the behavior of these energy systems in buildings, they are monitored using sensors and the data is stored for analysis in later stages. Nowadays a lot of data is recorded for different energy system components such as heating, ventilation and air-conditioning. The huge amount of data makes it unrealistic to manually analyze the data in detail. The simple most solution used by experts in the field is to use different types of visualization for analysis but still there is a chance of missing some interesting pattern of the energy system that might be helpful in detecting and diagnosing a fault. Therefore, it is required to develop such methods that can automatically detect various types of faults in the data.

Nowadays energy systems installations (e.g. solar cooling) devices have the problem, that the employed monitoring system along with existing methods for fault detection is not satisfactory. They are able to accurately detect few faults and are not able to detect and diagnose many fault scenarios correctly yet. One of the reasons might be the faulty sensors that records and transport wrong values. These wrong measurements will result the devices to shut down, although the system has no problem while in operation. On the other side, there can be situations where sensors report the measurements correctly, but the monitoring system is not able to detect these faults. In most such situations only the trained engineers can detect faults after a time consuming investigations of the relevant measurements. The circumstances can be troublesome specifically in case if dealing with solar based energy systems in buildings, as these systems are operational for a few hours a day and in case of fault the solar irradiation could not be used efficiently because of system faults. Therefore, it is required to detect faults and resolve it quickly for achieving good performance of the energy systems.

The first problem that is usually faced while monitoring the energy system in buildings is having a generic monitoring architecture for various types of energy systems, developed by different companies. The generic monitoring architecture need to handle the following issues

- Number and type of sensors required for evaluation of various energy systems and fault detection and diagnosis.

- The placement of each sensor for monitoring.
- There should be a naming convention for denoting the sensors, as this will help the experts in analysis of various types of energy systems from different companies.

Furthermore, the practical experience, of handling the historical monitoring data from sensors, has revealed that the data is mostly inaccurate and incomplete [8, 9, 10]. There are different issues observed with the data such as modifications in the configuration that are not correctly tracked, improper calibration problems of sensors, outliers in the data (data points that do not represent the behavior of the system), and some other issues in the data storage structure. Therefore, data sanitation is required for validating the data before the detailed analysis of the data. The data sanitation is the process of improving the quality of the data that can be further used in analysis. It has been observed that the conventional analysis method using the calculations found in standards such as EN 15316 produces unsatisfactory results if the quality of the data is low [11].

If the system has a duty cycle, there may be a variable that indicates the current state of the machine, e.g., a binary signal from the controller. In this case the duty cycle detection is straight forward by using this variable. Otherwise the duty cycle has to be derived from other process variables such as temperature levels or mass flows. For this detection a robust algorithm is required that can automatically process the data batches.The behavior of every module of the energy system in buildings varies between the operational and Off state of the system. The state information can be very helpful in finding the different patterns in the recorded data. Therefore, a method is required to automatically detect the operational duty cycle of machine from the recorded sensor parameters of the energy system, without involving any expert in the field.

The sensors are generally deployed in the fields to monitor and record the real life phenomenon such as temperatures, electricity consumption, in the energy systems of the buildings. Usually, these sensors in the field are operating in hostile environments; thus the sensor recorded data has high probability of anomalies existence in the data. The detection of such anomalies is necessary and is a challenging task due to the large amount of data recorded in the energy systems of the building. It is important to understand the different anomalies patterns and detect these patterns automatically for having a quality data for analysis and with less involvement of experts.

In order to evaluate the performance of various energy systems, it is required to have an automatic method for fault detection in the operational behavior of the energy systems. The behavior of the buildings energy system is quite complex to capture the faults pattern, therefore a method is needed for automatically detection of faults in the operation of systems without the help of experts. One of the easiest approaches, adopted by the experts in the field, is to use different visualization tools. But, the huge amount of data recorded during monitoring, makes it difficult for the experts to have a detailed performance analysis of the buildings manually or with simple visualizations. Moreover, there is a high chance for overseeing some important patterns in the data that may lead to faults in the different components of building, causing reduction in the energy efficiency. The methods that can automatically extract the different patterns in buildings data will help the experts to get more insight of the different parameters of the buildings energy usage and other processes such as factors causing faults in different components of the buildings.

1.3 Methodology

Data mining is one of the prominent research area that can help in automatically detecting the different patterns in buildings data; an example is clustering [12, 13, 14]. The process of

automatically finding various patterns in the data can make the analysis easy, feasible and less labor extensive [12, 13, 15]. After the detection of faults, in the operational behaviors of the energy systems, it is required to diagnose the detected faults. The diagnosis of faults will help to know the reason behind the detected fault. Subsequently, the maintenance operator can replace the required part of the energy system, saving cost and energy consumption. Thus, a method is required that can diagnose the faults using the detected fault information.

The focus of this research is to use data analytics and machine learning methods for developing a work flow for automatic detection and diagnosis of various anomalies in the recorded data. In addition, to find different fault patterns in the operation data of energy systems in the buildings, with minimum as possible support from the experts in the field. The proposed work-flow leads to novel contributions in the fault detection and diagnosis in the energy systems of building. Figure 1.6 is giving the overview of the methodology and shows the method against each problem statement.

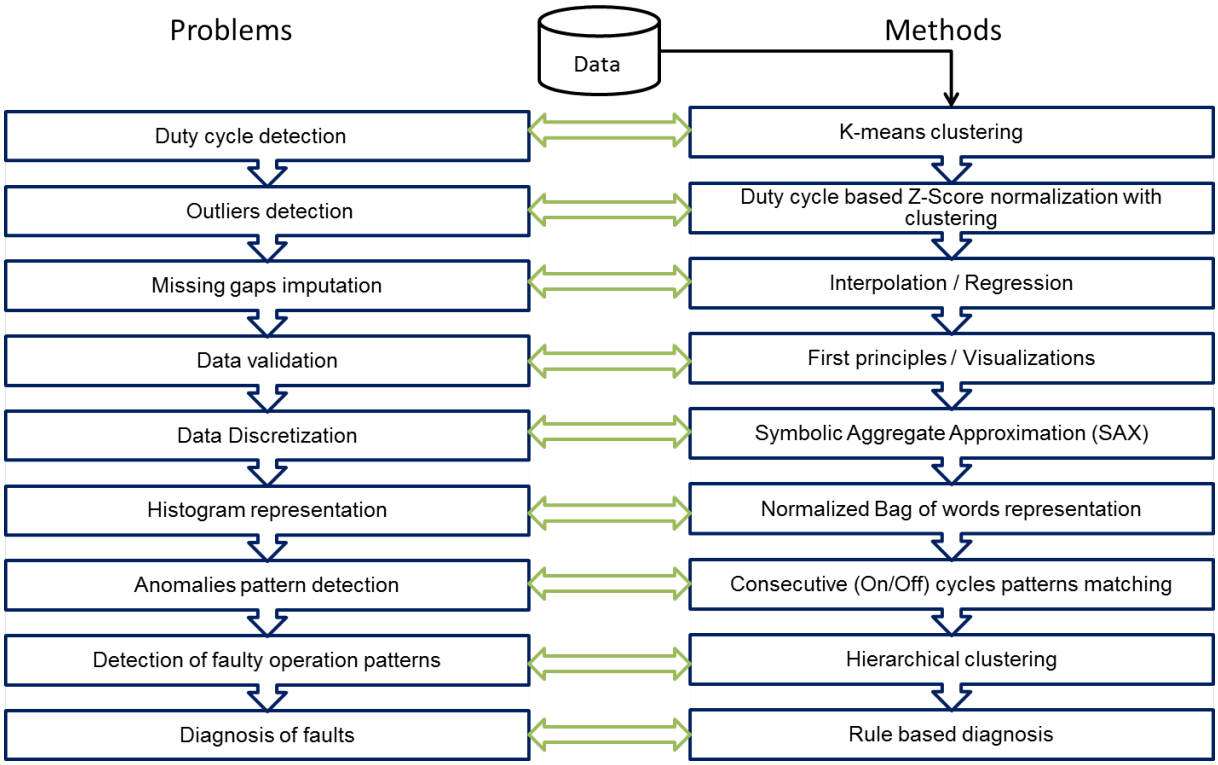


Figure 1.6: Overview of the methodology

The operational data of various energy systems have been used for the purpose of the validation of the proposed methodology such as solar based adsorption chillers, solar based absorption chillers, and air handling unit. The data was recorded and stored in a monitoring and analysis system JEVis. Furthermore, for analyses of the recorded data of the energy systems performance, a unified monitoring procedure developed in the framework of Task 38 "Solar Air-Conditioning and Refrigeration" initiated by Solar Heating and Cooling Programme Implementing Agreement of the International Energy Agency [16]. The unified monitoring procedure aims to generate set of performance figures of monitored SHC systems; estimate the primary energy savings of monitored SHC systems with the respect to a conventional system; and to enable the comparison between different monitored SHC systems.

In order to have a robust method for automatic detection of the duty(operational) cycle, the behavior of the different parameters such as temperatures, flows, energy readings and pressures of various energy system are analyzed. It is observed that the data has two different patterns showing the operational (On) and non-operational (Off) state of the energy systems. Therefore, on the basis of this assumption that the energy system behavior pattern will different in the two states, a method has been proposed using k-means clustering algorithm that clusters the data in two clusters representing Operational(On) and non-operational (Off) cycle.

As the behavior pattern of the energy system varies in two states i.e.operational (On) and non operational state (off), therefore this information is used for automatic detection of outliers. The outlier can be defined as those data point that do not correspond to the expected normal behavior of the energy system . So a method for automatic outlier's detection is proposed using Z-Score normalization based on the On/Off duty cycle state of the energy system, and subsequently, clustering the data by Expectation Maximization clustering algorithm [17]. The duty cycle based z-score normalization helps in highlighting the outlier's for clustering algorithm to cluster these points away from the normal behavior of the data. Thus helps in finding the invalid data points in the data.

Furthermore, the problem of gaps created in the data are handled using interpolation and regression methods, depending on the length of the gap. For short gap use of interpolation method is suggested, while for longer gaps use of regression is suggested. Moreover, validation of data is suggested before detailed analysis as it will save time rather than analyzing data that does not represent the behavior of the desired energy system. Therefore, first principles and visualizations are proposed to validate the data with first hand knowledge.

There are various types of anomalies in data of sensors recording, as sensors usually are deployed in fields for monitoring. A method has been proposed to automatically anomalies patterns in sensor data that is recorded in energy systems of the building. A crucial issue with monitoring data in energy systems is the data quality of the recorded data: due to problems in commissioning, transmission, storage or the sensor itself the data is often not feasible for further analysis, thus impairing the ability to detect faults in the operation of the energy systems. The focus lies on finding a method that covers all different possibilities of error patterns and has a high degree of automation so that it requires close-to-none configuration. The data is subjected to k-means clustering for the unsupervised classification of On/Off cycles and then transformed to a symbolic representation by applying the Symbolic Aggregate Approximation (SAX) method. The SAX symbols are converted to a Bag of Words Representation (BoWR) for each On/Off cycle. A pairwise comparison of the latter is carried out for the automated detection of anomalies patterns.

In order to automatically detect faults in the operation of the energy systems and to find various patterns in the energy system operation, a bag of words representation (BoWR) with subsequent hierarchical clustering has been proposed. The method uses the duty cycle detection information by the k-means clustering algorithm. The operational (On) cycles are of greater importance for finding the performance of energy systems and further in detecting and diagnosing faults. Additional features are suggested for achieving better results of clustering algorithm. The hierarchical clustering uses the BoWR of the On cycles of each feature for finding the various operational patterns of the energy system. The hierarchical clustering technique groups the data over a different scales by creating a cluster tree which is also called dendrogram. The dendogram shows a multilevel hierarchy of clusters, where the clusters (groups) at one level are joined together as clusters at the next level. This property of hierarchical clustering allows to decide the level of clustering that is most appropriate for the task it is used. In order to decide the best level

or optimum number of cluster, the gap statistics have been suggested. The cluster having the duty cycles with less average performance and longer duration is considered to be possible area of faults.

After clustering the data, additional information is added to each cluster such as average coefficient of performance (CoP) of the cluster and average time of On duty cycles in a cluster. Futhermore, additional parameters of operation have been integrated into the analysis, enabling to diagnose these faults after comprehensive discussion with experts in the field. An algorithm using various rules have been suggested for the diagnosis of faults in the operation of the energy systems.

2 State of the Art

This chapter discusses the state of the art concepts that are studied during the research. Buildings are import part of human life; as a major portion of human life is spend in these buildings. Therefore, the facilities for human comfort, provided in these buildings have always been considered as an important factor. In order to provide comfort in these buildings various types of heating, ventilation and air-conditioning (HVAC) systems are used in buildings, causing increase in the electricity consumption. Thus, for providing such facilities in buildings, it required to provide such solutions that are energy efficient, which will help in fulfilling the world energy requirements. Faults in these systems play an important part in decreasing the energy efficiency of the energy systems of the building. Therefore, an automatic detection and diagnosis of faults will help in improving the energy efficiency of energy systems in the buildings. There were various types of research topics found in literature that were in the direction of fault detection and diagnosis of the energy systems in buildings.

In the first section, the different types of Heating Ventilation and air-conditioning (HVAC) components are introduced. The details of chillers (adsorption/absorption) and various types of air handling units are discussed. In the subsequent section, the monitoring policy for solar heating and cooling (SHC) is elaborated, that is given in IEA SHC Task 38 monitoring policy for data recording of HVAC components. The use of IEA SHC task 38 helps in the analysis phase generalized for different architectures of the HVAC components, and also useful in comparing the various different types of HVAC systems in terms of performance. After discussing the HVAC components monitoring policy, the different types of faults and the methods used for detection and diagnosis of these faults are explained. The faults discussed include the errors due to the unavailability of data, errors due to data corruption, and faults in the operation of the energy systems. There are different types of solutions suggested by researchers e.g. rule based solutions using expert knowledge of the domain, and there solutions uses knowledge of other domains such as machine learning etc. Moreover, various methods for clustering time-series data are described that are used for finding the different patterns in the data. At the end different methods for finding the optimum number of clusters in data is discussed.

2.1 HVAC Components

This section will give an overview of the various HVAC components available and used in buildings. The detail discussion along with the theoretical architecture of various HVAC component used for

Heating, Cooling and Ventilation such as adsorption chiller, absorption chillers and air-handling unit is given in this section.

2.1.1 Heating and Cooling Components

A Heating, Ventilating, and Air-Conditioning (HVAC) system manages the devices (boilers, chillers, pumps, fans, etc.) for maintaining proper environment inside the buildings in a cost effective manner as can be noticed in Figure 2.1.

The following variables represent the proper environment in buildings [18]:

- Temperature : The temperatures between 68°F (20°C) and 75°F (25°C) are considered as comfortable for many people.
- Humidity: The humidity between 20% relative humidity and 60% relative humidity helps to keep the proper environment in buildings.
- Pressure: The rooms and buildings typically have a slightly positive pressure to reduce outside air infiltration. This helps in keeping the building clean.
- Ventilation: Indoor Air Quality is very important to keep the buildings environment comfortable for people. Therefore rooms and buildings have several complete air changes in the course of an hour.

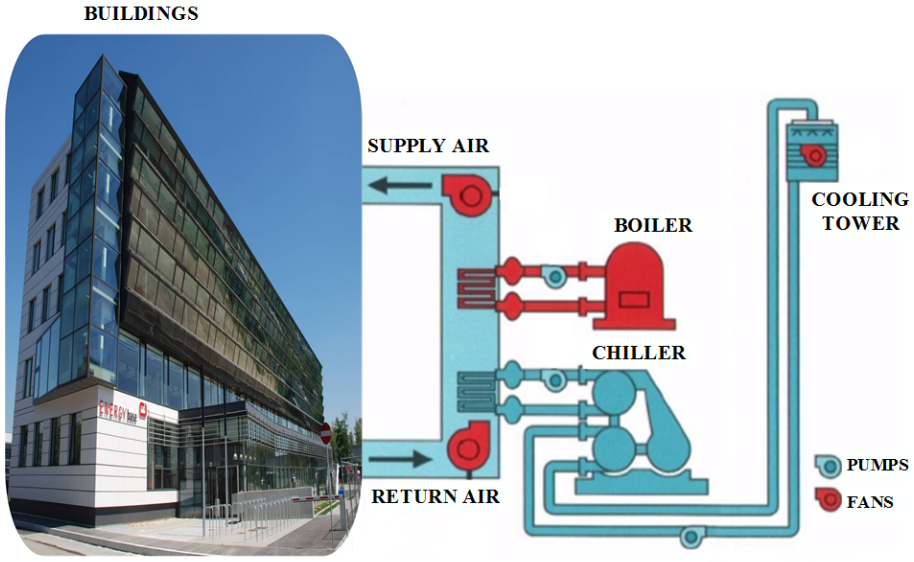


Figure 2.1: Basic HVAC (Heating, Ventilation and Air-Conditioning) System

In order to manage the HVAC components cost effectively, it is required to monitor these components and analyze them for better performance. These HVAC components are monitored by using sensor that senses the real life physical phenomenon like temperatures, pressures, Volume flow, and Electricity consumption by each component etc. Now days, we have a lot of data during the monitoring of the building HVAC equipment's. We record the data for long periods and manually analyze the data after certain period for getting some information regarding the performance of the building. As the data sizes are getting larger, it is difficult to analyze data manually using

different graphic tools. The benefit of gathering large data is lost if it cannot be analyzed for meaningful and interesting patterns. With the use of machine learning methods like clustering, hidden structure of the data can be discovered. These hidden meanings of the data can be useful in fault detection and diagnosis, which will further help in reducing energy cost.

There are different types of HVAC components used for heating, cooling and Ventilation in buildings. In some case one device provides all the facilities. Some of the common HVAC components are discussed as following:

2.1.2 Chillers

There are two kind of chillers available. The chillers are used in buildings for cooling purpose. The chillers produce the chilled water by using either a vapor-compression technique or absorption refrigeration cycle. The chilled water is then used to circulate through a heat exchanger for cooling and dehumidification of air and buildings as required. There are two kinds of chillers available that are discussed as following

2.1.2.1 Absorption Chillers

The fluid which is used in absorption chillers is a solution consisting of two liquids i.e. refrigerant and absorbent. As it can be observed in Figure 2.2, there are two parts which are connected to each other. The left part contains the liquid called refrigerant while the right part contains a solution of mixture of absorbent and refrigerant. The solution part will absorb the refrigerant vapor which will be generated by the left part causing pressure to reduce. This will cause the reduction of temperature at the refrigerant side, while the refrigerant vapors are being absorbed at the solution side. This process causes a refrigeration effect inside the refrigerant part. After absorbing the refrigerant vapors the solution on the right side of Figure 2.2(a) will become more dilute because of the higher content of refrigerant absorbed. This is called the "absorption process". The absorption process is usually an exothermic process; as a result, it must discard the heat out to the surrounding environment for maintaining its absorption capability [19].

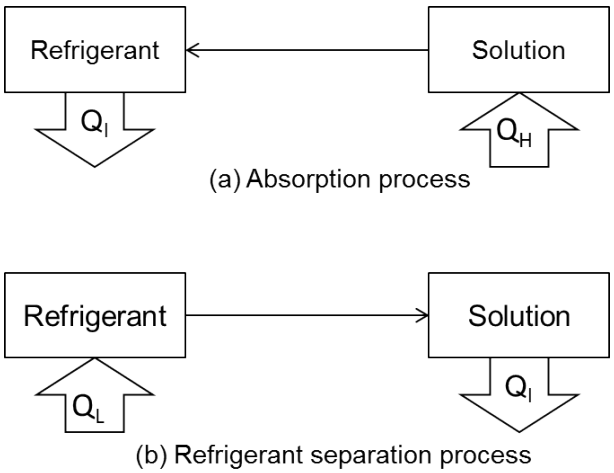


Figure 2.2: Absorption and Separation Processes

It is essential to separate the refrigerant from the diluted solution at the stage when the solution cannot further continue with the absorption process, because of the saturation of the refrigerant. At this point, the heat is normally used as the key for the separation process. As it can be observed in Figure 2.2(b) heat is applied to the solution part for separating the refrigerant from the solution. This will cause the refrigerant vapor to condense by transferring heat to the surrounding environment. So heat energy can be used for generating the refrigerant effect by using these processes.

In order to produce the cooling effect continuously it is required to have both the absorption and separation process run simultaneously. Hence, an absorption refrigeration cycle is a combination of both processes as can be seen in Figure 2.3. A circulation pump is required to circulate the solution as the separation process occurs at a higher pressure as compared to absorption process. The Coefficient of Performance (CoP) of absorption refrigeration can be obtained with the following Equation 2.1

$$CoP = \frac{Q_7}{Q_{6a} + W_{pump}} \tag{2.1}$$

Where Q_7 is the cooling capacity obtained at evaporator, Q_{6a} represents the heat Input at the generator, whereas, the W_{pump} is the work input for the pump. The W_{pump} is usually neglected for the analysis purpose; therefore the equation for coefficient of performance can be written as Equation 2.2

$$CoP = \frac{Q_7}{Q_{6a}} \tag{2.2}$$

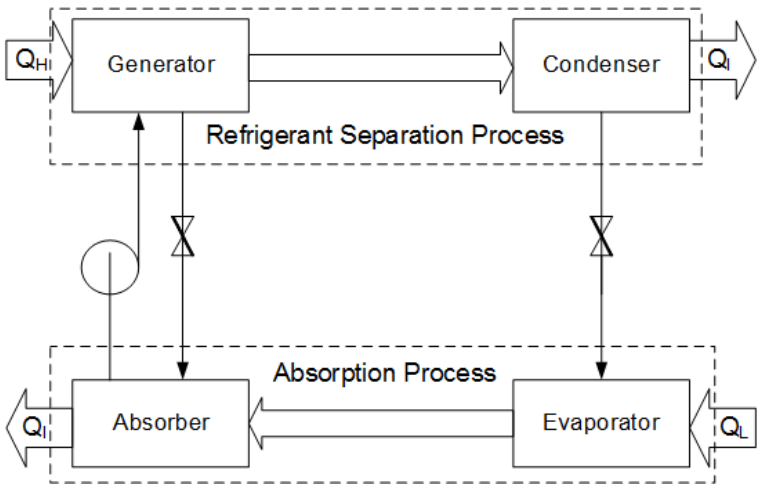


Figure 2.3: Continuous Absorption Process

The Performance of absorption refrigeration systems is heavily dependent on the following [20, 21]

- The chemical and thermodynamic properties of the working fluid (refrigerant and absorbent).
- During the liquid phase, the refrigerant and absorbent should have a margin of miscibility within the operating temperature range of the cycle.

- It is highly recommended to consider the difference in boiling point between the pure refrigerant and the mixture at the same pressure. A better Coefficient of performance will be achieved with larger difference in boiling points.
- In order to maintain a low circulation rate between the generator and the absorber, it is required to a select a refrigerant with high heat of vaporization and high concentration within the absorbent.
- It is beneficial to consider the transport properties that can influence the heat and mass transfer, e.g., viscosity, thermal conductivity, and diffusion coefficient.
- The liquids (refrigerant and absorbent) must be non-corrosive, ecofriendly and should be available at low-cost.

2.1.2.2 Adsorption Chillers

The second type of chillers is adsorption chiller that uses water as refrigerant. The structure of an adsorption chiller consists of 3 different chambers i.e. evaporation, middle chamber (receiver, generator) and condenser, as can be seen in Figure 2.4. The two parts in the middle chamber works as receiver and generator for a specific time period. After the specified time period each chamber replaces the responsibility of each other, that is simply done by applying hot water to one part for specific period of time and to the other part in the next turn.

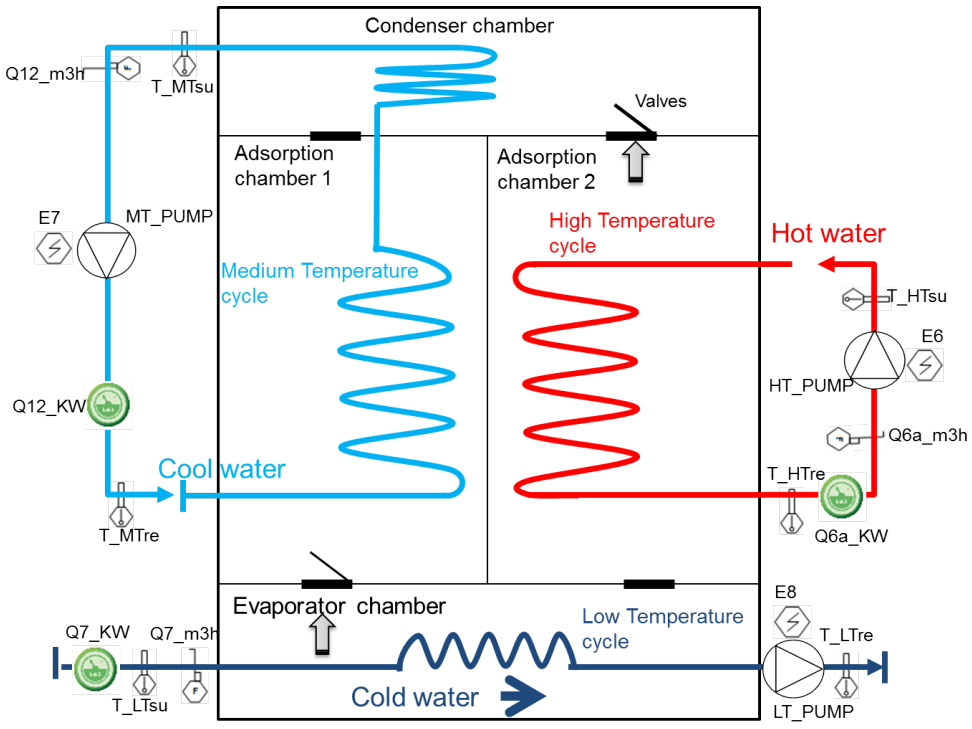


Figure 2.4: Adsoption Chillers

The working of adsorption chiller can be defined in the following steps [22]

- The water is evaporated in the lower chamber called evaporator which makes the water cool in Low Temperature (LT) cycle.

- The evaporated water is adsorbed on the receiver side that is the middle chamber using silica-gel.
- The adsorbed water is then deadsorbed with the heat provided from the Hot water cycle.
- The deadsorbed water is condensed and brought back to the evaporator.

The different temperatures ($T\_HT_{su}, T\_HT_{re}, T\_MT_{su}, T\_MT_{re}, T\_LT_{su}, T\_LT_{re}$), volume flow ($Q6a\_m3h, Q7\_m3h, Q12\_m3h$), energy meters ($Q6a\_KW, Q7\_KW, Q12\_KW$) and electricity consumption meter ($E6, E7, E8$) sensors can be observed in Figure 2.4. The pumps in the heating (HT_Pump), cooling (LT_Pump) and heat rejection (MT_Pump) cycle circulates the water in the respective cycles. The placement of the sensors is done according to the IEA Task 38 for recording the data, which can be used for analysis of the system. The use of IEA Task 38 will help to generalize the monitoring policy of different chiller architecture.

2.1.3 Air-Handling Unit

Air ventilating unit or air handling unit is a part of building HVAC system that is responsible for the exchange or replacement of air inside the building, in order to provide high indoor air quality. The indoor air quality includes many factors such as indoor temperature, humidity, odors, smoke, oxygen replenishment, and controlling carbon dioxide. The process of ventilation contains circulation of air i.e. exchange of air to the outside as well as pumping back air within the building. Ventilation is important part of the HVAC system of the buildings as it maintains the acceptable indoor air quality inside the buildings. The ventilation systems can be divided into two methods i.e.

Natural type: This is a simple most ventilation system in building where the air exchange is done without using fans or any other mechanical device. It operates through windows or louvers in case if the spaces are small. The concept used in such system is the temperature of air e.g. the warm air is allowed to rise and flow outside through the opening that will cause the cool air outside to enter into lower part of the building. These systems use very less amount of energy, but do not ensure comfort especially in warm or humid climates.

Mechanical type: The mechanical ventilation system provides the facility of controlled quality of indoor air by using various handlers. Such systems control the excess of humidity and odors via dilution or exchange of air with outside air. Though, in excess humid environments, it will require much more energy to remove the extra moisture from the ventilation air. Nowadays these kinds of systems can be found in the kitchens and bathrooms and use mechanical exhausts to control odors and humidity. The factor involved in such systems is the flow rate such speed and size of the fan and noise level.

There are more complex systems available for air handling unit that include several additional features involving temperature control (heating or cooling the space), humidity etc. Usually, these systems require individual control of multi-zones inside one building e.g., shopping malls, offices, hotels, hospitals, and schools. Furthermore, these systems can be categorized with a single duct or double duct system. Another classification can also be used i.e. constant or variable air volume for multi-zone systems. The single duct are more energy efficient as compared to double duct system, however it has been noticed that the variable volume systems are more efficient. The benefit of such system is that, the heat recovery systems can easily be incorporated in the main

air-conditioning units. The drawback of such systems is the additional duct space and operation costs [23].

The thermal distribution system using water can also be used to heat or cool a space in buildings. These kind of systems used the direct heat transfer process between water and the indoor air in the building. There are various types of terminal units that are commonly used such as chilled ceilings and chilled floors, and chilled-water systems containing fan coils. The space required for thermal distribution system in water systems is significantly less space and it also can provide the functionality of individual room control. Additionally, simultaneous cooling and heating in such systems is also possible, nevertheless, the maintenance cost of such systems is high [23].

One of most commonly used ventilation system is the refrigerant based system. The cooling process in such systems is done in a short distance from the installed unit, and they are mostly either fixed in windows or mounted on wall. There are various types of systems that come under this category such as splits, and packaged air conditioners. These systems are simple and provide room control at low cost, and lesser initial cost as compared to the central systems in buildings. The drawback with systems is the low flexibility in terms of air flow rate, condenser and evaporator sizes, and most importantly the higher price of power consumption per kW as compared to central systems [23]. The complex air handling unit can do the functionality that includes air conditioning, distribution systems with air-handling equipment, and lastly the air/liquid distribution systems that controls the suitable interaction between the air-handling unit and the building. There are various arrangement settings available for air handling equipment and distribution systems [24]. For maintaining appropriate relation of air conditioning and distribution, an efficient design of the duct systems is required such as single duct or double duct, depending on the air flow control strategy e.g. constant air volume (CAV) or variable air volume (VAV).

The single duct provides the functionality of heated and cooled air using a single duct. These type of systems can be further classified depending on the air flow control strategy. Figure 2.5 shows the typical single duct of type single zone and constant volume system configuration. As compared

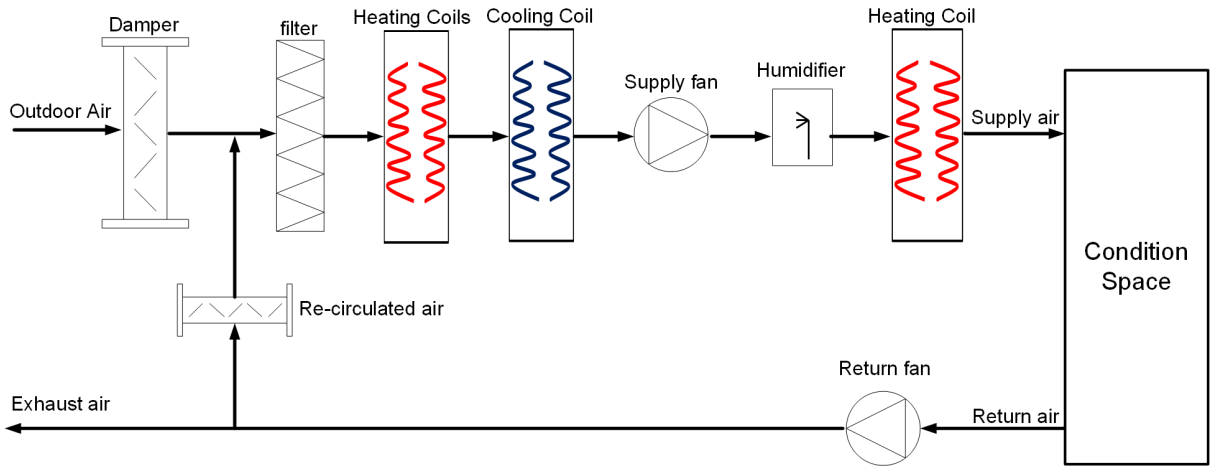


Figure 2.5: Configuration of single duct, single zone and constant air volume system

to single duct, the dual-duct arrangement the air is conditioned in the central equipment, while the distribution of air is done using two parallel ducts. The ducts are used to carry the cold and warm air separately in two ducts. Figure 2.6 is denoting the dual-duct with single-fan and constant air volume[24].

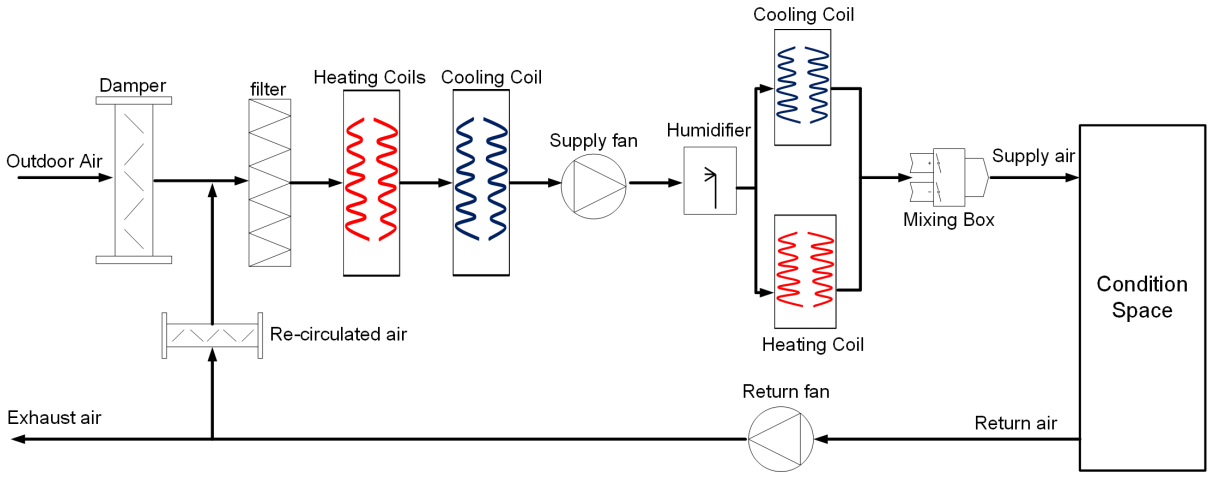


Figure 2.6: Configuration of dual duct, single zone and constant air volume system

2.2 Data Monitoring Procedure

The data acquisition and monitoring procedure are an important issues for analyzing the data in detail. Figure 2.7 represents the general architecture for the data acquisition process. The data can be recorded from different energy systems of building, situated at different locations. The data is fetched after some time from the controller recorded in the energy system and the feteched data is then stored in a monitoring and analysis system called such as OpenJEVis [25]. OpenJEVis is an open source solution providing functionality for monitoring and analyzing the energy data. It provides data storage feature and different visualizations of data. Additionally it also provides a meta-data structure for tagging different states of data like faulty etc. The benefit of using sophisticated data management system is that the data is recorded directly into the storage with additional information. A pool of FDD algorithms can be applied on the data saved in the database, and relevant meta information can be saved in it. The saved meta data can be useful in data analysis of the different types of energy systems in the buildings. The

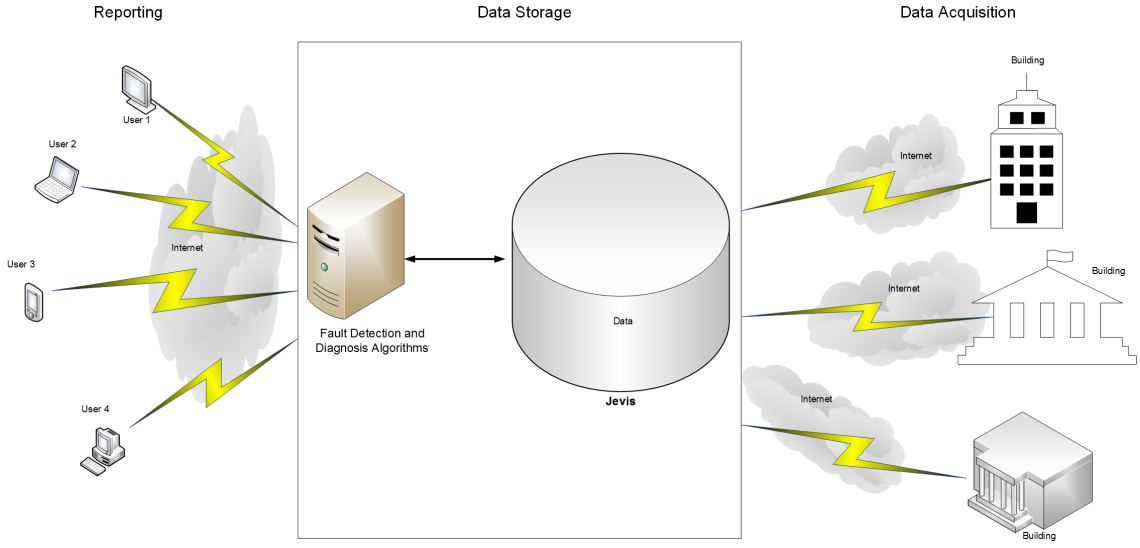


Figure 2.7: General Monitoring Framework for Data Acquisition

relevant information can be viewed by user like facility manager without the knowledge of the whole process, and will help them in decision making regarding the energy system performance.

The International energy agency (IEA) has launched a technology initiative called "IEA Solar Heating and Cooling Programme (SHC)". There are various research issues addressed in this implementing agreement, and one of the topic that is discussed is IEA SHC Task 38 "Solar Air-Conditioning and Refrigeration". The IEA SHC Task 38 (subtask A3a-B3b:"Monitoring procedure for solar cooling system") defines a generic monitoring policy that provides information on sensor locations and naming for the evaluation of systems, evaluation of the system performance, and comparison of different energy systems [16]. The idea behind having a common monitoring procedure is to help in evaluating the monitored systems and compare them with a particular conventional system. The decision of selecting any conventional system for reference always depends on the fact that whether the monitored system is based on SHDC (Solar Heat Driven Cooling) or DEC (Desiccant Evaporative Cooling). Therefore, IEA SHC Task 38 [16] has described a standard diagram as can be seen as Figure 2.8, in order to enable a clear and homogeneous representation of monitored systems and of the conventional system to be taken as reference for comparisons. There are different modules added in reference diagram, but in case if some modules are missing then the Figure 2.8 can be redrawn accordingly.

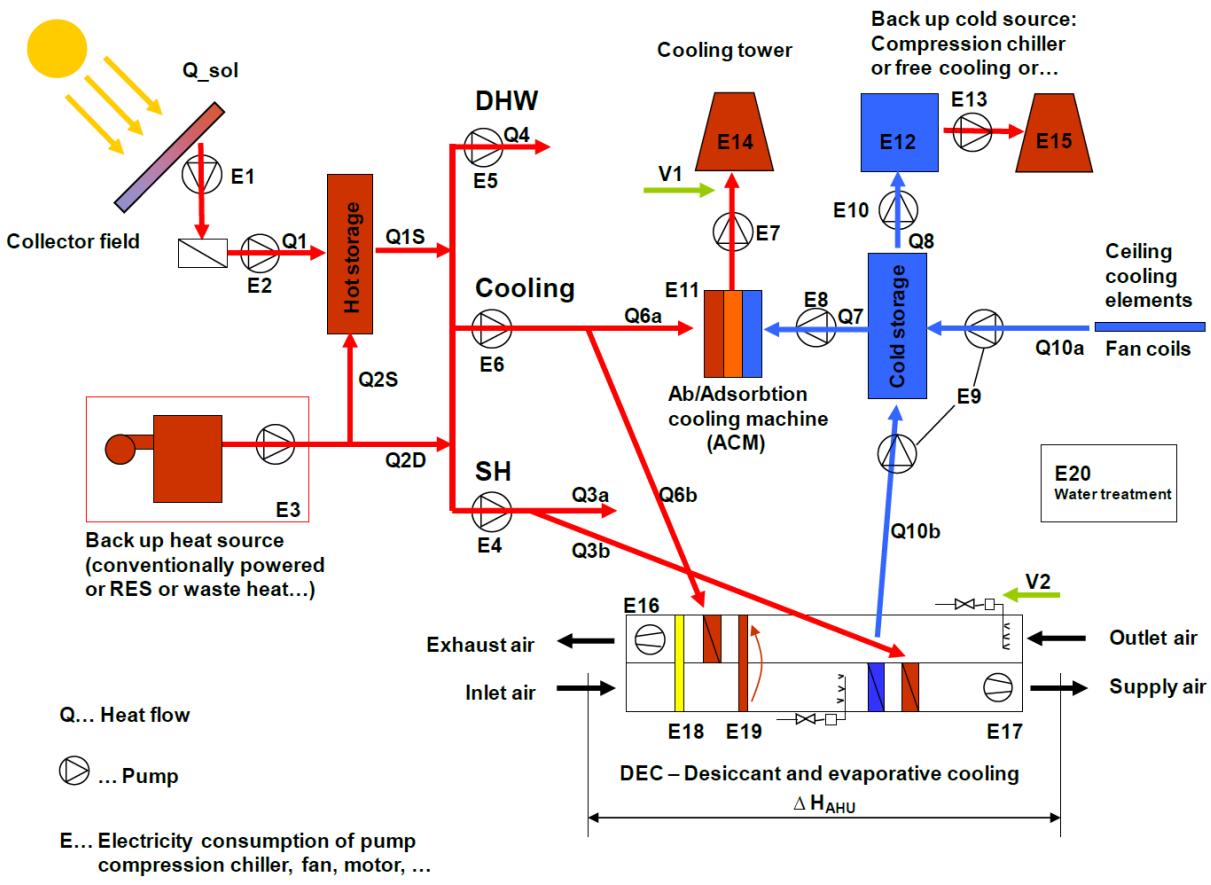


Figure 2.8: Reference Diagram for Solar Heating and Cooling System [16]

IEA SHC Task 38 also suggests the necessary measurements for the evaluation of a system with a certain energy performance indicators. The variables used in the reference diagram are explained in the tables given below.

Table 2.1 shows the electric parameters required for comparison of systems and finding the performance indicators of each system.

Electricity consumer [kWh]	**Label**
Heating System	
pump collector field (primary loop)	E1
pump collector field (secondary loop)	E2
pump boiler hot-storage (including internal boiler consumption)	E3
pump hot-storage to space heating (SH)	E4
pump hot-storage to domestic hot water (DHW)	E5
Cooling System	
pump hot-storage to cooling machine	E6
pump cooling machine (ACM) to cooling tower	E7
pump cooling machine (ACM) to cold-storage	E8
pump cold storage to cold distribution	E9
pump back up source - cold storage	E10
absorption/adsorption cooling machine (ACM)	E11
compression chiller (back-up system)	E12
pump compression chiller to fan (back-up system)	E13
fan, cooling tower	E14
fan of compression chiller (back-up system)	E15
Desiccant cooling/ dehumidification System	
fan exhaust air	E16
fan supply air	E17
motor for desiccant wheel	E18
motor for heat recovery wheel	E19
Water treatment System	
water treatment for wet cooling tower and humidifier for DEC	E20

Table 2.1: Required Electrical Parameters [16]

Moreover, Table 2.2 describes the water consumption parameters required for finding the performance parameters of the systems that can further help in comparison of different systems.

Water Consumption [Liter]	**Label**
water consumption for wet cooling tower	V1
water consumption for DEC humidifier	V2

Table 2.2: Required Water Consumption parameters [16]

Furthermore, Table 2.3 discusses the required thermal parameters that can used for finding the performance parameters of any system.

Thermal Energies [kWh]	Label
solar irradiation on total collector aperture area	Q_sol
solar thermal output to hot storage	Q1
heat output from hot storage	Q1S
boiler thermal output (fossil) into storage	Q2S_fossil
renewable energy source (RES) thermal output into storage	Q2S_RES
fossil boiler thermal input (fossil) bypassing hot storage (directly used)	Q2D_fossil
renewable heat source (RES) thermal input bypassing hot storage (directly used)	Q2D_RES
space heating (SH) consumption (conventional)	Q3a
space heating (SH) consumption (ventilation system)	Q3b
domestic hot water (DHW) consumption	Q4
hot storage input to cooling machine (ACM)	Q6a
hot storage input to DEC-system (sorption regeneration)	Q6b
cold output ACM to cold-storage	Q7
cold output back-up chiller or free cooling to cold-storage	Q8
cold storage output to cold-distribution	Q10a
cold storage to Air Handling Unit (AHU)	Q10b
Enthalpy difference - Air Handling Unit(Inlet Air =>Supply Air)	ΔH_{AHU}

Table 2.3: Required Thermal Parameters [16]

As usually such comparison of systems is based on many assumptions of the conventional system which can lead to misleading results. To tackle these issues IEA SHC Task 38 has developed a common procedure for evaluation of the performance of monitored solar heating and cooling installations. The development of such standard procedure will enable to even comparison the results of several systems as well. The details of the evaluation of the performance can have the following three levels.

- Basic level: In the first level the system overall performance and the share of the renewable for providing cooling facility to the building are analyzed. It includes basic Information on COP, along with the Primary Energy Ratio, and Costs for providing the facility for a period of monthly evaluation as well as yearly evaluation. The following data will be required that are related to basic level are as following:
  - The electric meters reading, that shows the overall electricity consumption
  - The thermal energy reading on the cooling side of the chiller.
  - The last information it will need is the solar radiation.

  The major output of that can obtained at the first level can be summarized as following:
  - It will provide the ratio between the cooling load and the total electrical consumption that excludes the distribution.
  - The ratio between the cooling load and the total electrical consumption that includes the distribution as well.
  - The Primary energy ratio of the monitored system presenting the contribution of the Renewable Energy Systems.

  - The Primary energy ratio for the monitored system showing the Renewable Energy System that contributes as fossil.
  - The Primary energy ratio for the conventional system in order to set a reference for comparisons

- Medium level: It contains the details of basic level of monitoring procedure with additionally focusing on the evaluation the solar heat management efficiency. The following information is required for medium level monitoring:

  - The Heat flux meters for measuring the energy transfer between the different components.
  - It needs electricity consumption metering, or, it requires a rough calculation with control input parameters, or operational period information such as operation hours and average values of performance of each component

  The medium level provides the following:

  - the efficiency of the Storage.
  - System efficiency.
  - Efficiency of the Solar heat management.
  - Unused Solar energy.
  - Total heat load.

- Advanced level: The Advanced monitoring procedure gives the details of the evaluation of the primary energy savings. Furthermore, it also explain the details of any specific COP‘s of components and groups. This level of evaluation procedure are complex and costly as having such details need large number of sensors. The different types of output that can be generated at this level are:

  - Fractional primary energy saving of the monitored system calculated for renewable energy systems or fossil. This is used for comparing with a conventional system. There are 5 different version calculated for DEC systems.
  - Fractional primary energy saving of the monitored system of renewable energy systems, as compared with a conventional system.
  - Fractional solar consumption of the available solar energy and the overall heat load.
  - Water consumption
  - Water treatment electricity consumption.
  - Electrical coefficient of performance (CoP) of the chiller.
  - Electrical coefficient of performance (CoP) of the heat driven cooling system.
  - Electrical coefficient of performance (CoP) of the heat driven cooling system along with the solar collectors.
  - Thermal coefficient of performance (CoP) of the heat driven cooling system.

2.3 Faults Detection and Diagnosis in Buildings

The FDD (Fault detection and diagnosis) research for HVAC systems started in the late 1980‘s [26] as there were methods explored for automated FDD in vapor compression based refrigeration [27]. In the early 1990‘s, new ideas and research for FDD applications in building systems were developed and tested in laboratories. The focus of most of these investigations were to find automatic methods for FDD vapor compression equipment e.g. refrigerators, air conditioners, heat pumps, and chillers. The methods for FDD generally used the phenomenon of measuring the temperature and/or pressure at various locations in a system for a valid thermodynamic relationships in order to detect and diagnose common faults. The International Energy Agency (IEA) has commissioned the Annex 25 collaborative research project in early 90‘s, which has done initial work on real time simulation of HVAC systems for building processes optimization, and FDD [28]. The study has also identified some common faults for various types of HVAC systems, and investigated many methods for faults detection and diagnosis. The IEA had moved forward by conducting another study in order to validate and demonstrate the investigated FDD systems in real buildings [29].At the same time, the U.S. Department for Energy (DOE) had developed a diagnostic tool for the whole building. The diagnostic tool focus was on detecting anomalies in the whole building and its energy consumption, while it also focused on FDD for outdoor air ventilation and economizing [30]. ASHRAE Technical Committee on Smart Building Systems had also become active in the mid to late 1990s for FDD research, and had sponsored many research projects [31, 32, 33].

There are various types of faults that can be observed during the operation of any energy based system. Therefore, each type of fault requires to be handled with different method of Fault detection and diagnosis. The different types of faults in HVAC systems can be categorized as following

2.3.1 Unavailability of the Data

The first common category of faults is the unavailability of data during the operation of system, due to uncontrollable reasons such as calibration, communication link damage or sensor damage etc. The missing data causes greater influence on the analysis of HVAC system as there will be no information available [34, 35, 36]. Secondly, the other issue will be how to fill these gaps that can be a proximate to the real value of the system because different data can cause absolutely different results. Therefore, handling the missing data problem is very tricky and it is also important to detect these missing gaps in the data, as this is a significant factor to indicate the reliability of the data. These missing gaps cannot be ignored as it will be lowering the amount of meaningful calculations. It is very critical to identify and fill these missing gaps, as data with fewer gaps represents the good quality of data [37]. There are numerous missing values pattern types [34, 38] such as

- Missing completely at random (MCAR): The pattern of the missing values in a data set is called as missing completely at random (MCAR) if the events occur entirely at random and is not dependent on any other observations.
- Missing at random (MAR): The pattern of the missing values in missing at random (MAR) type is that the missing values of an instance occur randomly but is dependent on other observed variables. This means that the data can be retrieved using the information of other variables.

- Missing not at random (MNAR): In this case, the missing value occur but they are neither MCAR nor MAR.

There are different methods available in literature to encounter the missing values. The selection of solution depends on the nature of the data and other parameters like computations, precision, robustness, accuracy etc. [39]. One of the simplest way of handling missing values is to ignore the instances having missing data in case when the percentage of missing data is too small or the amount of remaining data is large enough for analysis [36]. There are two major categories of data imputation techniques [40, 41, 42]: Single Imputation (SI) and Multiple Imputation (MI). The single imputation method gives a single value to be inserted in the missing place, whereas, multiple imputation offer several imputations (the average of all the values can be selected for inserting in the missing data place). Some of the single imputation methods include the following techniques

- **List wise deletion**: it is a simple imputation method where the data containing missing gaps are deleted [34].
- **Pairwise deletion**: In this scenario only those records are deleted where the required variables for analysis are missing.
- **Hot-Deck Imputations**: It is an imputation method for where each missing value is replaced with an already observed value from a similar observation [43, 36].
- **Global Imputation on missing attributes**: In this imputation technique the information of frequency or central tendency is used to fill the missing values [36]. The example of such method is replacing the missing values with either the mode, mean or median of the data [44].
- **Global Imputation on non-missing attributes**: In this method the information of the correlation between the missing data and observed data is used. Regression is one of the common methods used in this kind of techniques [45, 44, 46, 47], other than interpolation [48]. Furthermore, neighbor based methods also fall in this category where missing gaps are filled with the nearest value in terms of distance such as K-nearest neighbor [49, 50]. Moreover, co-clustering technique also known as two way clustering can be used for replacing missing values [36].

2.3.2 Corrupt Data

The second common problem that is faced during monitoring of HVAC system is getting incorrect data measure by sensors. As these sensors record different physical real life phenomenon e.g. temperature, pressure and electricity consumption of HVAC components in buildings online, and are located in the field, thus there is always a chance of getting the data corrupt. The incorrect data can lead to wrong results. Therefore it is required to diagnose such problems and validate the data before detailed analysis. The common methods used to tackle such issues are to check status of sensor, physical range check, detection of gaps and constant values, tolerance band method, material redundancy detection, signal gradient test, extreme value check, physical and mathematical based models and data mining techniques [51, 10, 52].The data can be labeled as A, B and C on the basis of data validation techniques where A represents correct, B represents doubtful and C represents wrong data [52]. There is also a possibility of assigning a confidence value between 0 and 100, in order to provide refined information about data [37].

2.3.2.1 Sensor faults pattern detection in data

The sensors are deployed to monitor and record the real life phenomenon such as temperatures, electricity consumption, etc. in energy efficient buildings. As these sensors in the field are generally operating in the non-controlled environment, therefore the sensor fault probability is high. At the same time, the large amount of data recorded in the energy efficient buildings has made it a challenging task to detect such faults manually. It is important to understand the different sensor faults patterns and detect these sensor faults automatically for having a quality data for analysis.

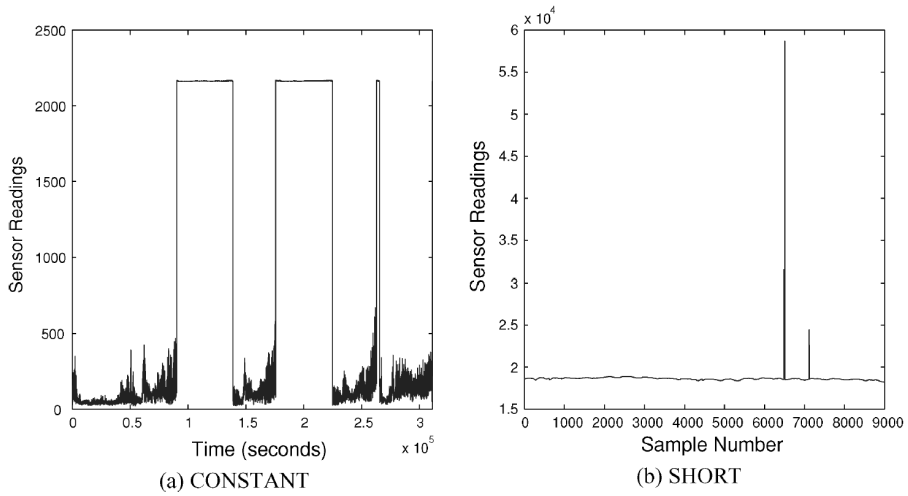
The data analysis has become a vital process for extracting the meaningful scientific information from the collected data [53, 54, 55]. It is essential to ensure the quality of the data before it is used for the detailed analysis. There are already a number of sensor deployments where the faulty sensors readings can be observed. The reason for such sensors faults are either the incorrect hardware design, improper calibration or low battery levels [54, 56, 57, 58]. One of the major causes for sensor faults in data is calibration problems [59, 60, 56]. The calibration errors in sensors can deviate the measurements from real behavior in different ways [54]:

- **Offset fault**: the deviation of the measured value from the real value with a constant amount.
- **Gain fault**: the rate of the recorded data differ from the expected rate.
- **Drift fault**: the change in different parameters assigned to sensor's original calibration formulas during the deployment.

There are different solution suggested in literature [61, 62] using the spatial correlation across sensor nodes for developing online sensor calibration methods. The classification of these faults can be done by using the statistical pattern recognition [63, 64, 65].

The different patterns of sensor faults are found in the recorded sensor data while in the faulty state by [54]. The different fault patterns found are

- **Short faults**: These are the singleton outliers or the abrupt change in the successive data, as can be seen in Figure 2.9a.



(a) CONSTANT (b) SHORT

Figure 2.9: Fault Patterns. a) Constant Faults. b) Short Faults [54]

- **Constant faults**: In this type of sensor faults the value of the sensor is stuck at one value such as in Figure 2.9b.
- **Noise faults**: The normal sensor data with suffcient voltage can be seen in Figure 2.10a, whereas, 2.10b shows the Noise sensor fault due to low voltage.

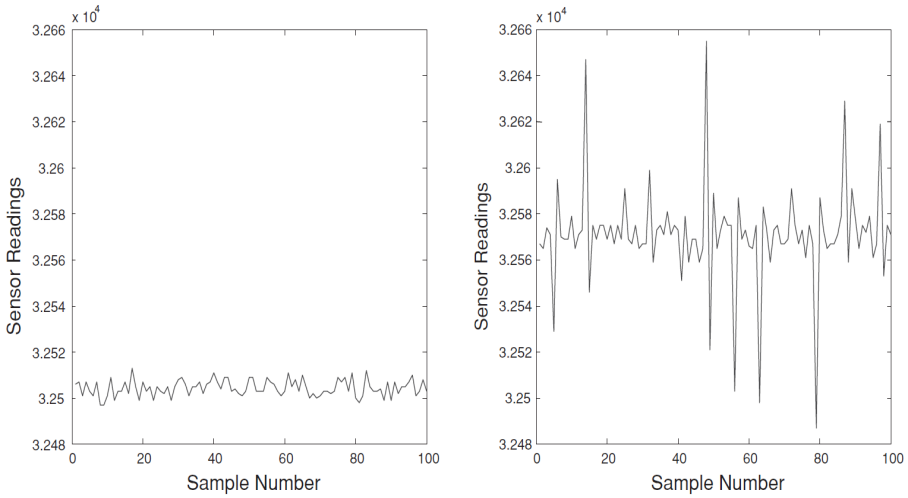


Figure 2.10: Fault Patterns. a) Normal Data. b) Noise Fault [54]

Such faults cause deviation from the normal patterns demonstrated by a real sensor readings [54, 56]. These fault patterns have been observed in numerous real-life deployments [53, 59]. Thus, it is important to understand these patterns and develop methods for automatically detecting them in the data.There are different methods available in literature for the detection of such sensors faults.

Rule based methods

In rule based method the standard deviation of sample readings within a window N is calculated. There are certain rules created for detecting sensor faults with standard deviation information of the window size N and threshold. For example, in case of Constant sensor fault the standard deviation will be zero. The performance of rule based methods depends on the window size N and the threshold and for the selection of best value for the window size N, requires the domain knowledge [54]. In order to automate the process of selection of threshold, the histogram method has been proposed in [54, 59]. In histogram method the time series data is divided into groups of N samples. The histogram of the standard deviations for these groups of N samples is created. For selection good threshold value it is required that the histogram have a (distinct) mode. Still the threshold value selection depends on the value of windows size [54].

Time Series Analysis methods

The real life phenomenon such temperature, ambient light, etc., normally demonstrate a repeating pattern (diurnal pattern). The observation of such sensor data like temperature for longer periods of time can be used to detect these diurnal pattern and time scale temporal correlations. The details of these temporal correlations can be exploited for constructing the model for sensor data using time series analysis [66].

Autoregressive integrated moving average (ARIMA) is multiplicative seasonal model that can be used to model time series data for sensor fault detection [54]. The multiplicative seasonal model is one of the commonly used methods for modeling and estimating the data having periodicity feature [67]. The following equation shows the model as,

$$S_t = S_{(t-1)} + S_{(t-p)} - S_{(t-p-1)} + N_t - \theta.N_{(t-1)} - \Theta.N_{(t-p)} + \theta.\Theta.N_{(t-s-1)}, \tag{2.3}$$

Here the parameter S_t is the predicted sensor value at time t, and p determines the periodic behavior of timer series in the sensor data; such as temperature values can exhibit similarities with period of 24 hours [54]. N_t is the noise process at time t. the two parameters, θ and Θ, are calculated using the training data of the sensor that can be used for modeling. The value of S_t depends on the value of the sensor values at time $t - p - 1$. The detail explanation of seasonal models is given in [67].

The sensor data faults can be detected using the different length of look ahead (L) steps. The predicted values are compared with the real data and decision can be made on the difference between these two values for declaring the data as faulty [54]. They have tested with different values of look ahead (L) starting from one. It is noteworthy to note that with L=1 the model has performed better for Short sensor faults.

Learning Model Based Methods

The learning model based methods are more suitable for predicting values that may not be spatio temporally correlated. There are different learning based models currently used in research for detection of sensor faults, such as Hidden Markov Models (HMMs) and neural networks [54]. In learning based models the model are created with training data and later the model can be used for sensor fault detection. One of the most important learning model used is artificial neural networks (ANN) [68, 69, 70]. The values predicted by the learning based models are compared with the real value. These values are compared and if the difference between the values is greater than certain threshold, in such case the data is marked as sensor fault. [54, 68].

2.3.3 Faults in the Operation of the Energy Systems

The FDD has shown improvement in three areas of building engineering i.e. Commissioning, operation and maintenance. The commissioning stage involves the initial installment of devices and is usually the faults are caused by incorrect installation (e.g., fans installed backward), incorrectly sized equipment, and improperly implemented controls etc. These problems are usually diagnosed by visual inspect and certain manual tests. There can be the option of saving the data loggers for later detailed analysis of the behavior of the systems. The data may not be available in the commissioning stage as there will be no data for analysis, but in later stages it can be very helpful in the operation phase for diagnosing faults, and at the end for maintenance. The steps as shown in the Figure 2.11 below can be automated and recorded for having a complete functional test of system and should be done and recorded during the life cycle of the device [71, 72]. FDD in operation and maintenance of any engineered system can be seen in Figure 2.11

The basic components of fault detection and diagnosis (FDD) systems consist of techniques for detecting faults and subsequently diagnosing their causes. There are various methods used in the field for detecting and diagnosing different faults. These techniques vary due to the fact that how they use the knowledge for formulating the diagnostics. The information can be used as priori knowledge such as first principle based model methods or data driven based methods like black

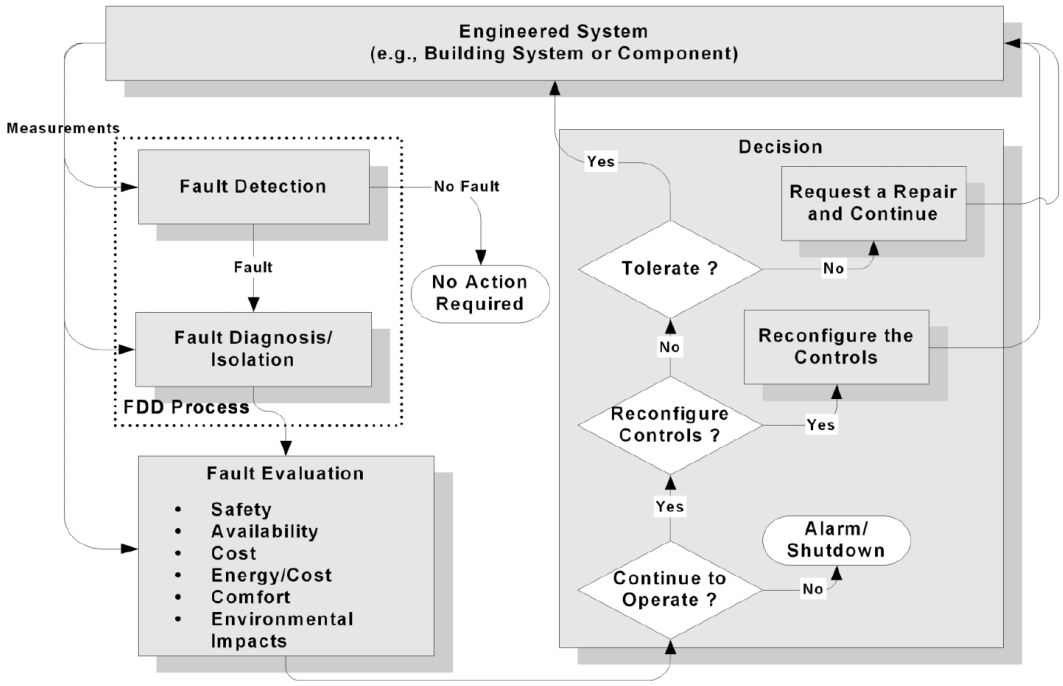


Figure 2.11: Process of fault detection and diagnosis [51]

box models. In first principle based model methods, the priori knowledge is used to generate a model that will be used as a reference model and the difference between the actual readings and from the reference model will be the basis for the identification and evaluation of faults [73, 74]. Whereas, the data driven based techniques generates a model from the behavior of the actual data of the process and does not depend on the priori knowledge [75].

The model based methods can be further categorized as quantitative models and qualitative models. The quantitative models rely on the quantitative mathematical relationships depending on the underlying knowledge of the physics of the processes. The quantitative model methods include those approaches that create the model based on the detailed physical properties, or those depending on the physical processes. In comparison the approaches in the qualitative model include rule based or models depending on the qualitative physics. The rule based systems can be either based on expert rules called as expert systems or simple limit checks [73, 74, 51].

In comparison to the methods depending on a priori knowledge of the process, the process history based methods rely solely on the process history, thus it requires a huge amount of historical data for generating a model. One of such methods includes black-box methods, where models are created purely from the data. The other method that comes under this category may include gray box models where first principles or engineering knowledge is used to identify the mathematical form of terms in the model. For gray box models the parameters (e.g. coefficients in the model) are calculated from the process data. There are different statistical derived methods included in black box models such as regression, artificial neural networks (ANN), and pattern recognition techniques [75, 51].

The authors in[51] has given a broader classification of FDD methods that can be seen in Figure 2.12 summaries the different types of solutions available for the FDD in HVAC system.

The quantitative models are based on physical and engineering principles; therefore provide the

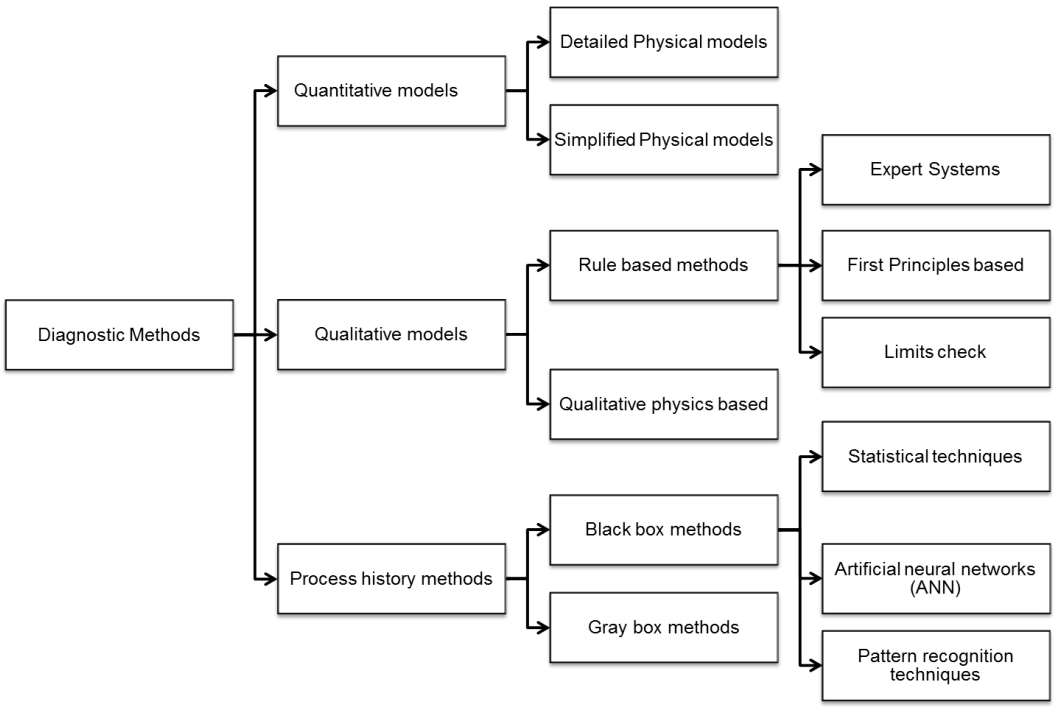


Figure 2.12: Different methods for fault detection and diagnosis

most accurate predictions of the output and can also be used to model the transient phase of the system as well. The details provided by the model makes easy to differentiate between the normal and faulty operation, but at the same time needs greater effort for modelling, complex and are computational intensive. The faults during transient state are mostly detected by using the dynamic physical models. The list of different available dynamic models for vapor compression can be found in [76]. The authors in [76] also had discussed a dynamic centrifugal chiller model using first principles for fault detection and diagnosis (FDD). The model developed by [76] can be used as basis for finding FDD as most of the other vapor compression models requires detailed physical characteristics for the system and subsystems, which is not always easy to achieve and apply [51]. The dynamic physical modelling for FDD implementations can be found in many research papers [31, 77, 78, 79]. In comparison to the detailed physical modeling techniques, the simplified physical modeling methods are less computationally expensive.

The quantitative modeling is another approach that is heavily depending on the priori knowledge of the system. In these models the knowledge of the system is expressed in terms of qualitative relationships or knowledge to draw the conclusions concerning the state (faulty or non-faulty) of the system and its components. Expert system is a type of qualitative models where the models are generated using the deriving knowledge statements, obtained from the process history data. Normally in most of the expert systems, the human experience with a process is used to predict the state of the machine as proper or faulty operation.

In order to make the buildings energy efficient, the models of buildings are developed and simulated for energy performance. For better designing and able to handle the dynamic nature of the different properties of building, each component of the building can be modeled as an active part, thus different component of the building will make a complex network [80]. There are many

energy modeling methods that are generally used for predicting the buildings performance during the design phase. The actual energy consumption reading usually deviates from the predicted value during the modeling phase [81, 82]. Some of the reasons for this deviation are the dynamic parameters such as occupants behavior in the building, climate, and buildings properties [82]. The agent based methods can be used to handle such dynamic parameters by using agent based modeling, e.g. agent based modeling is used for handling the dynamic nature of occupants behavior with the impact on energy consumption in commercial buildings [81, 82], and house hold impact of energy consumption managing ventilation system in residential buildings [83]. There are several bottom up models developed for the agent based modeling and the most common issues with such methods are the handling complexity and uncertainty problems. The authors in [84] have proposed a framework using pattern oriented modeling for agent based modeling to handle the complexity and uncertainty problems.

The qualitative models techniques can be further partitioned into two types i.e. rule based models and qualitative physics based models as can be seen in Figure 2.12.

The two qualitative modeling techniques work with the fundamental knowledge of the process or system to diagnose faults. There are other qualitative based models such as bond graphs. The bond graph represent a graph where the edges are denoting the energy equations, whereas the nodes are either symbolizing system elements (e.g. resistance, capacitance etc.) or sources [85]. The case based reasoning [86]method and abstraction hierarchies based on decomposition method [74], are some solutions for qualitative modeling found in the literature.

The rule based modeling methods come under qualitative method, using a priori knowledge to develop a set of rules in style of if-then-else rules, and the inference mechanism searches the rule-space to predict conclusions. The rule based systems can either use the knowledge of the experts or first principles for deriving rules. There are also methods available where first principles are used for making a rule based models [30, 87, 88]. In [89], the authors have proposed a FDD method for cooling equipment using the difference between the measured values and the readings of the model using thermodynamic state values at the steady state of the rooftop air conditioners for the detection and diagnoses of faults. It has been recommended to use at most, nine temperatures sensor values and one relative humidity sensor data. The suggested values are

- Ambient temperature of the air at the condenser input (T_{amb}).
- Return air temperature at the evaporator (T_{ra}).
- Return air Relative humidity (F_{ra}) at the evaporator coil or wet bulb temperature (T_{wb}).
- Evaporating temperature for the system (T_{evap}).
- Suction line superheat (T_{sh}).
- Condensing temperature (T_{cond}).
- Liquid line subcooling (T_{sc}).
- Compressor outlet temperature (T_{hg}).
- Air temperature rise across the condenser (ΔT_{ca}).
- Air temperature drop across the evaporator (ΔT_{ea}).

The steady state model has been used to define the association between the actual values and the expected output states of the model in a normally operating system. The residuals (difference between the actual value and expected value) are the basis used by the detection classifier to detect fault and further diagnosis can done to find the reason of faulty behavior. The authors proposed a method for FDD in [89], where the detection of fault is estimated on the basis of the probability that that the current behavior is normal. The state of the machine is considered as a faulty in case the probability decreases below the threshold, which is known as the fault detection threshold. The rule based classifier is used to perform fault diagnosis as can be seen in Table 2.4. The set of rules for each fault are dependent on the difference between the reference model and actual reading of each measurement, while the fault occurs. These set of rules for diagnostic can be seen in Table 2.4 for the five different faults and seven output measurements established by [90]. The table shows each measurement increases (↑) or decreases (↓) in response to a specific fault at steady state condition. The following are the five types of faults diagnosed are,

- **Refrigerant Leak**: A refrigerant is a mixture used for cooling in different devices. The refrigerant phase change from liquid to gas and vice versa cause cooling in air conditioning systems. The air conditioner having low refrigerant is either due to undercharged or may be due to leakage. The performance and efficiency of the air conditioner is at maximum when the refrigerant charge matches the manufacturer's specification. Furthermore, a refrigerant leak is also harmful to the environment.

- **Compressor Valve Leakage**: The compressor valve is important for keeping the pressure difference during the phase change of the refrigerant. Some of the reason for compressor's valves becoming inefficient includes valve warping from overheating, lack of lubrication, and different deposits on them preventing them from sealing properly such as carbon. In case of compressor valve leakage, the efficiency of cooling will be affected.

- **Liquid Line Restriction**: A restricted liquid line will cause starvation in the evaporator, causing low evaporator pressures. Furthermore, the starved evaporator will cause starvation in the compressor and condenser. This will cause low pressure in both the condenser and evaporator.

Table 2.4: Rules for FDD [90]

Fault Type	T_{evap}	T_{sh}	T_{cond}	T_{sc}	T_{hg}	ΔT_{ca}	ΔT_{ea}
Refrigerant Leak	↓	↑	↓	↓	↑	↓	↓
Compressor Valve Leakage	↑	↓	↓	↓	↑	↓	↓
Liquid-Line Restriction	↓	↑	↓	↑	↑	↓	↓
Condenser Fouling	↑	↓	↑	↓	↑	↑	↓
Evaporator Fouling	↓	↓	↓	↓	↓	↓	↑

- **Condenser Fouling**: The main function of the condenser is to remove the unwanted heat from the system to the environment. One of the main reasons for low efficiency in

cooling systems is condenser fouling that is usually caused by crystallization that causes duct blockage.

- **Evaporator Fouling**: Evaporator is the part of the cooling system where the phase change of the refrigerant occurs that causes cooling in the system. One of the main reasons for evaporator fouling is crystallization.

It can be noticed that each of the faults results in a different grouping of increasing or decreasing measurements against their normal values. The rules discussed in Table 2.4 are efficient fault models and can be considered as a generic model for roof top air conditioners.

The authors in [7] has done extensive experimental for evaluation of performance of the FDD technique proposed by Rossi and Braun [89]. The authors have used simple rooftop air conditioner in laboratory for test and have observed it for steady state and transient phase for various fault levels. In order to determine the statistical threshold for the detection of faults, the authors have used the data without faults for training the models for normal operation, while in case of transient state the data with faults have been used for evaluation of the FDD performance. The sensitivity of faults detected and diagnosed can be seen in Table 2.5. The Table 2.5 shows each fault detected level with one data point represented with "First" and at the level when all steady state points ("All") demonstrate the fault from the database of transient test. The Table 2.5 shows the five faults along with the respective loss in capacity in percentage , COP lost in percentage, the change in the superheat (ΔT_{sh}), and the change in temperature of sub cooling (ΔT_{hg}) at these detectable levels. These results demonstrate that most of the faults can be detected and diagnosed earlier then the decrease in capacity of 5% is attained. The results further elaborates that the technique is less sensitive to compressor valve leakage and evaporator fouling in terms of COP.

Table 2.5: Performance Evaluation of FDD [7]

Performance Index	**Refrigerant Leakage**		**Liquid Line Restriction**		**Compressor Valve Leak**		**Condenser Fouling**		**Evaporator Fouling**	
	First	All	First	All	First	All	First	All	First	All
Fault Level (%)	5.4	Max	2.1	4.1	3.6	7.0	11.2	17.4	9.7	20.3
Loss Capacity (%)	3.4	>8	1.8	3.4	3.7	7.3	2.5	3.5	5.4	11.5
Loss COP (%)	2.8	>4.6	1.3	2.5	3.9	7.9	3.4	5.1	4.9	10.3
$\Delta T_{sh}(^\circ F)$	5.4	>11	2.3	4.8	-1.8	-3.6	-0.6	-1.6	-1.7	-2.7
$\Delta T_{hg}(^\circ F)$	4.8	>10	2.4	4.8	0.0	0.0	1.8	2.3	-1.2	-2.7

The researchers have proposed a fault detection and diagnosis method for chillers where they have done the detection and diagnostic in one step [91]. The proposed FDD was developed using the data from an operating chiller, while considering a multivariate linear regression method as a reference model to estimate values for process variables values of non-faulty state of chiller. Later these estimated values from the models are used to generate residuals, which are the difference between the recorded value and the value from the reference model. The chiller is categorized into seven parts for identifying the fault modes, which are condenser, liquid line immediately with the condenser that includes a filter drier, compressor, evaporator, expansion valve, liquid line with solenoid and sight glass between the other liquid line, and the crankcase heater. The authors had developed 58 different fault modes. The symptom for any faults is given as the difference between the measured and derived variable from the model made with normal operation data. The symptom patterns are organized and the identical faults modes were removed, making the

fault modes from 58 to 37. Furthermore, different experiments were performed and patterns of these residuals are compared to faulty condition patterns. In order to diagnose the fault mode a comparison is done with the corresponding set of symptom patterns. The scores are calculated to indicate the probability of the symptom patterns matching each fault mode. Each fault mode symptoms are compared to the corresponding measured symptom and score of 0 to 10 is assigned. If the measure symptom exactly matches the symptoms of fault mode then the fault mode is assigned high score i.e. 10. Similarly with less matching the fault mode is assigned less score.

The Fault modes with high scores were considered as existing faults in the chiller, whereas fault modes with low scores are neglected, and faults with medium scores were considered as possibly existing. In order to do the different test for the fault modes twenty different measurements were used that includes compressor oil level, temperatures, power consumption, and pressures along with some derived variables, such as liquid sub cooling, pressure drop and superheat. The Table 2.6, shows the symptom patterns associated with a particular fault in each row against the twenty sensor variables. The symptom patterns were denoted by the arrow pointing up (↑) indicating that the observed value for the variable is greater than of the reference model. Similarly, an arrow pointing down (↓) shows the corresponding symptom to the observed value for the variable is less than the reference model, and a horizontal arrow (→), specifies that the fault has no effect on the corresponding variable.

Table 2.6: FDD symptom patterns [91]

Fault Modes	**Compressor Suction Pressure**	**Compressor Suction Temperature**	**Compressor Discharge Pressure**	**Compressor Discharge Temperature**	**Compressor Pressure Ratio**	**Oil Pressure**	**Oil Temperature**	**Oil Level**	**Crankcase Pressure**	**Compressor Electric Power**	**Subcooling of refrigerant**	**ΔT Refrigerant and cooling water**	**ΔT Cooling water**	**Inlet Temperature at Expansion Valve**	**Filter Pressure drop**	**Evaporator Outlet Pressure**	**Super heat**	**ΔT chilled Water**	**Evaporator Outlet Temperature**	**Number of Acting Cylinders**
Compressor, Suction side, increase in flow resistance	↑	→	→	→	→	↓	→	→	↓	↓	→	→	→	→	→	↑	↑	↓	→	→
Compressor, discharge side, increase in flow resistance	↑	→	↑	→	→	↑	→	→	↑	→	→	→	→	→	→	↑	↑	↓	→	→
Condensor, cooling water side, increase in flow resistance	→	→	↑	→	→	→	→	→	→	↑	↓	→	↑	→	→	→	→	→	→	→
Fluid Line increase in flow resistance	→	→	→	→	→	→	→	→	→	↓	→	→	→	↓	→	→	↑	↑	→	→
Expansion valve, control unit, power element loose from Pipe	↑	→	→	→	→	↑	→	→	↑	↑	→	→	→	→	→	↑	↓	↑	→	→
Evaporator, Chilled Water Side, Increase in flow resistance	↓	→	→	→	→	↓	→	→	↓	↓	→	→	→	→	→	↓	↑	↑	→	→

The authors in [92] have proposed FDD method using the chiller model developed by Gordon and Ng [93] and the pattern matching technique suggested by researchers in [91] and discussed in

Table 2.6, as a module in their solution. The proposed method used the thermodynamic model for fault detection, while used pattern recognition for fault diagnosis for selected faults using the expert knowledge. A set of 17 different measurements including flow rates, pressures, and temperatures were used to detect 4 different faults i.e. refrigerant line flow restriction, refrigerant leak, evaporator water-side flow resistance, and condenser water side flow resistance. Furthermore, FDD system is partitioned into three sections i.e. one part is used to detect faults during the transient phase; second part was dealing with faults during the chiller steady state (operational); and third section was detecting faults when chiller is not operational (Off).

The chiller off status module was used to detect faults, mostly the temperature sensors faults where the sensor measurements are compared with other sensor readings. It is expected that the readings will reach to one value when the chiller is shut down and sensor readings attain the steady state. After this, the comparison information can be used during commissioning such as for managing calibration faults during commissioning. The transient state module is used for the first 15 minutes after the chiller is operational. The faults during transient phases are detected by comparing the measurement trends of variables during the transient phase and baseline trend of the normal start-up of the chiller. Similarly, the steady state (operational) module was used after the chiller reaches steady state. The steady state module was used to evaluate the performance of the system and detect and diagnoses the required faults. The evaluation of the performance was done by using the thermodynamic models suggested in [93]. The linear regression was used to predict the values and can be used as reference model for comparison with actual measurements. The Table 2.7 shows the predefined patterns corresponding to the various faults using a rule base method. Moreover, the authors in [94] modified the method by replacing the rule based with statistical pattern as a reference model, but the drawback with statistical pattern method is the availability of training data for normal and faulty operation, which is usually difficult to implement in the field.

Table 2.7: FDD pattern proposed in [92]

Fault	**Discharge Temperature**	**High Pressure Liquid Line Temperature**	**Discharge Pressure**	**Low Pressure Liquid line temperature**	**Suction line temperature**	**Suction Pressure**	ΔT_{cond}	ΔT_{Evap}
Restriction in refrigerant line	↑	↓	↓	↓	↑	↓	↓	↑
Refrigerant leak	↑	↓	↓	↓	↑	↓	↓	↑
Restriction in cooling water	↑	↑	↑	↓	↓	↓	↑	↓
Restriction in chilled water	↑	↓	↓	↓	↓	↓	↓	↓

In contrast, the expert systems are computer based applications emulating the decision making ability of a human experts. The expert systems are developed from the knowledge of domain experts by taking interviews, and converting the collected information in the form of if-then-else statement and saving it into a database [71]. There are different prototype systems developed to

perform diagnoses of operational problems in commercial building systems such as in [95], where the authors have developed an embedded expert system for monitoring packaged HVAC equipment. The expert system for FDD in building systems have been suggested by the researchers in [96]. Other than this, there are papers available in literature where the comprehensive description of use of expert systems in industries is provided [75, 97].

The qualitative physics based models that has been placed under qualitative model in the Figure 2.12, predicts the state of a system with incomplete or ambiguous knowledge of the physical process [98]. For predicting the state it uses two basic techniques. The first technique is to derive the qualitative confluence equations from the ordinary differential equations that are demonstrating the behavior of the processes under observation. These derived equations can be solved with incomplete information using a qualitative algebra. In the second method ordinary differential equations can be derived for qualitative behavior that demonstrates the physical behavior of the system. The obtained qualitative behaviors can be used as knowledge for FDD [74]. The applications of qualitative physical models for FDD can be observed in [99] using qualitative simulation or by using the qualitative process theory (QPT) [74]. The main benefit of using qualitative physics based models is that they enable to predict the information about the process without having the complete details.

One of the most widely used methods is data driven models where both inputs and outputs are known and measured in advance. The basic concept of the data driven model is to relate the measured inputs to outputs mathematically. The transformation from input to output can be done with various techniques and then using it as a priori knowledge in a diagnostic system. This transformation procedure is also called as feature extraction. The process based models are more appropriate for those problems where the theoretical models of behavior of the system are poorly developed. They are more suitable for those systems where the training amount of data is available. The black box method is one of such methods where the model features or parameters have no physical significance. The black box approaches include various techniques such as regression techniques, fuzzy logic, artificial neural networks (ANNs) [100], principle component analysis, partial least squares, pattern recognition methods, and clustering techniques etc. There are a lot of researchers working with black box approach for FDD [32, 33, 101, 102, 89, 14, 103, 104]. The black box models are the simplest models to develop and do not need the understanding of the underlying process physics of the system.

The two methods were used to detect 8 different types of faults in air handling unit (AHU) during laboratory experiments [105]. The first method used the differences between the measured and predicted variables for the detection of faults. The predicted values were generated with the normal operating conditions. While in the second approach, in order to detect faults, a comparison between the estimated parameters using autoregressive moving average with exogenous input (ARMX) and ARX models and the normal parameter is done. The different types of detected faults in AHU included the failure of the chilled-water circulation pump, failure of the supply and return fans, stuck cooling-coil valve, complete failure of temperature sensors, failure of the supply and return air fan flow stations, and complete failure of the static pressure sensor.

Furthermore the faults described previously in [105] has been detected using artificial neural network (ANN) [106]. The training of the ANN was done with the normal data, as well as the data that characterized the eight faults detected in [105]. There were seven normalized residual inputs to the ANN, and the output consists of nine values pattern representing the eight faults and the normal state of the AHU as can be seen in Table 2.8. The Table 2.8 shows the symptom residual, where $+1$ is representing the dominant residual having positive value, -1 for the negative

residual; while the remaining residuals are assigned 0. The patterns shown in Table 2.8 are used for training the ANN method.

Table 2.8: FDD patterns for ANN [106]

Fault diagnosis	Supply pressure	Difference in supply and return airflow	Supply air temperature	Control signal to cooling coil	Supply fan speed	Return fan speed	Cooling coil valve position	Network Outputs
No fault	0	0	0	0	0	0	0	1 0 0 0 0 0 0 0 0 0
Supply fan	-1	-1	0	1	-1	0	0	0 1 0 0 0 0 0 0 0 0
Return fan	0	1	0	0	0	-1	0	0 0 1 0 0 0 0 0 0 0
Pump	0	0	0	1	0	0	0	0 0 0 1 0 0 0 0 0
Cooling coil valve	0	0	0	0	0	0	1	0 0 0 0 1 0 0 0 0
Temperature sensor	0	0	-1	-1	0	0	0	0 0 0 0 0 1 0 0 0
Pressure transducer	-1	0	0	0	0	0	0	0 0 0 0 0 0 1 0 0
Supply fan flow station	0	-1	0	0	0	0	0	0 0 0 0 0 0 0 1 0
Return fan flow station	0	1	0	0	0	0	0	0 0 0 0 0 0 0 0 1

The ANN method proposed in [106] was extended by authors in [107].The authors proposed two ANN models for detection and diagnosis of air handling unit (AHU) faults [107]. The AHU was divided in to various modules i.e. the flow-control subsystem, the cooling-coil subsystem, pressure control subsystem, and the mixing-damper subsystem. The task of the first ANN was to detect the subsystem showing faulty operation, while the second ANN was assigned the task to diagnose the cause of the fault.

There are methods available in literature where the model parameters depend on the first principles and have physical significance. Such methods come under the category of gray-box models or also known as mechanistic models. The linear or multiple linear regressions are mostly used by gray box models to estimate model parameters from measured inputs and outputs, at the same time preserving the physical significance of the parameters for model. There are various studies available in literature where parameter estimation is applied to building systems [93, 108, 109]. Another gray box model known as the characteristic parameter model for chillers has been proposed in [110]. This model comprise of five sub models, where each models is denoting components or characteristics of the system. The suggested five sub models rely on the first principles of thermodynamics and are given as following:

- Coefficient of performance (COP).
- Compressor.
- Condenser.
- Evaporator.

- Expansion valve.

The International energy agency (IEA) has launched a technology initiative called "IEA Solar Heating and Cooling Programme (SHC)". There are various research issues addressed in this initiative. IEA SHC Task 48 (subtask: B6)"Report for self-detection on monitoring procedure" categorizes the current faults scenarios and their possible solution that can further help in the development of a systematic error detection system [111]. For each type of error, based of these categories, different error detection methods can be developed. Following are the steps for achieving the goals:

- Different system errors collection occurred in different demonstration sites.
- System error characterization.
- Methods development for common system errors detected.
- Describing the minimum additional required sensors for improved fault detection. methods.

The IEA SHC Task 48 has classified the different errors in five groups:

- **Systematic errors**: These errors are triggered by having a poor design and causes system failures or mall functions. The examples and methods are discussed in Table 2.9 given below:

Table 2.9: Systematic errors and solutions

Examples of the errors	Methods used for detection
• The design fault in the hydraulic system or during the installation. • Improper installation of the valves. • Incorrect selection of the dimension of components such as pumps. • Fault due to the wrong positioning of the components. • Fault due to the closed valves.	In such faults a detailed analysis of the system has been proposed and it has been concluded that automatic detection of these faults is a difficult process.

- **Sensor errors**: These errors are caused either by sensor or cabling defects. Additionally, bad installation, wrong placement can also cause these errors. The example and possible solutions are discussed in the Table 2.10.

Table 2.10: Sensor errors and solutions

Examples of the errors	Methods used for detection
• Sensor problems due to bad cabling. • Sensors anomalies such as data drift and offset. • Sensor calibration issues. • Problems due to bad parameterization.	• Range value check for each sensor. • Range check can be system specific as well. • Cross checking with other sensors values. • Analyze the time difference between two consecutive time stamps along with the values recorded at these time stamps.

- **Wrong Control Strategy**: These error are caused because of the wrong implementation of the system control. There might be several reasons for such errors such as programming errors in the control code, inappropriate control strategy, sensor errors causing wrong control decision, or by cabling errors. The examples and possible solutions are given in the Table 2.11

Table 2.11: Control strategy errors and solutions

Examples of the errors	Methods used for detection
• Errors due to bad controller parameterization. • Errors due to the wrong placement of the sensor causing wrong controller signal. • Errors due to wrong control code in various different conditions. • Errors due to wrong strategy such as control of fan speed control for heat rejection not available.	• Detect with performance indicators such as CoP. • Check with experience and certain rules. • Detailed analysis of data for a few slots of operations timing. • Comparison with the results of the simulation of the system. • Step response of the controller during the commissioning phase.

- **Component defects**: These errors are are caused by damage of a single component such pumps, motors, etc. It results in a complete shutdown of the system. The examples and solutions are discussed in Table 2.12

Table 2.12: Component defect errors and solutions

Examples of the errors	Methods used for detection
• Error due to the defective motor of valves. • Error due to the defective pumps. • Errors due to blockage of the valves.	• Analysis of the points showing the drop in the pressures. • Analysis of the points showing the drop in the pressures. • Comparison with the simulators results. • Testing with the performance indicators. • Comparison with the efficiency curves of the components.

- **Performance degradation**: These errors are caused due to the aging factor of the com-

Table 2.13: Performance degradation errors and solutions

Examples of the errors	Methods used for detection
• Errors due to Fouling. • Errors due to valve blockage or partly blockage. • Errors due to aging of components. • Errors due to blockage of filters.	• They are considered as the difficult errors to be detecteds. • Compare with the historic data. • Comparison with the simulators results. • Testing with the performance indicators. • Comparison with the efficiency curves of the components. • Experience or rule based checks.

ponents and results in low performance of the systems such as fouling effects. Table 2.13 gives the examples and solutions of the errors.

Furthermore, machine learning techniques can also be used for detection and diagnosis of faults in buildings. There are different faults that can be detected in buildings using the information from the installed electricity consumption meters with the machine learning algorithms [112, 113]. Moreover, the prediction of electricity consumption for each HVAC component can be done using various available parameters. In order to find such parameters, the multivariate analysis can be used for calculating these parameters and this will help the building operators to manage the devices, thus helping to save energy [114]. The machine learning techniques can also be used for extracting abnormal behavior information from the data such as clustering. The clustering technique was used for finding the similar daily performance patterns in the buildings [12, 115], detecting the abnormal performance with electricity consumption [116], and for enhancing the performance optimization algorithms [117]. At a larger scale, wavelet transformations with clustering was used for the classification of electrical demand profiles of buildings [118].

2.4 Clustering Time Series Data

Clustering is one of the methods of data mining extensively used in the research, where all similar objects are grouped in to relevant or same groups without having the advanced knowledge of the details of each group. Clusters can also be explained as all those elements make a cluster grouping that have maximum similarity with other elements within the group, but have minimum similarity with the objects in other groups. The clustering is a suitable approach in a scenario where the grouping of unlabeled is required for data analysis as it identifies structure in an unlabeled dataset by organizing data into similar groups. Figure 2.13 shows the clustering of same types of elements.

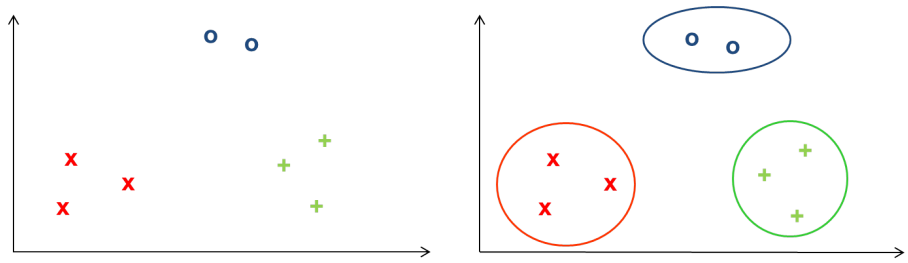


Figure 2.13: Clustering of different types of objects

The advancement in sensor technology and the increase in the power of data storage and processors, gives the chance to real world applications such as buildings monitoring system to store and keep data for a long time. The huge amount of time series data has given the chance to analyze the time series data for many researchers in data mining field. Therefore, many researches and projects focused in analyzing time-series in different fields for various reasons such as sequence matching, abnormality detection, motif discovery [119, 12], for indexing, classification, clustering [120], segmentation [121], visualization [122], detecting patterns, trend analysis [123]. Furthermore, there are other various research projects going on that have focused to improve the existing techniques [124, 125]. There are various surveys and reviews regarding the time series clustering and have focused on the comparative aspects of time series clustering experiments [15, 126, 127, 128, 129, 130]

Time series clustering is the special type of clustering. A series having nominal symbols from a specific alphabet is generally known as temporal sequence, whereas, a sequence of continuous real values, is called as a time series [131]. A time-series data can be categorized as dynamic data because of the change in the feature values is a function of time. In order words, the value of each point of a time series has one or more observations that are made chronologically. Time series data is a kind of temporal data that usually have high dimensional and large in size data [130].

The behavior of the buildings operation is quite complex, so clustering can be used to discover interesting patterns in such complex operation patterns that are usually saved as time-series datasets. There are several research challenges in clustering such complex systems such as methods required to recognize dynamic changes in time-series data. The following points highlight the importance for clustering time-series datasets,

- Clustering can be used to discover interesting patterns in the data as time series databases contain valuable information.
- As normally it is difficult to handle the time series databases manually due to large size of data. Therefore, it is preferred to deal with organized datasets rather than huge datasets. So, time series data can symbolized as a set of groups of similar time-series data in clusters or with a hierarchy of abstract concepts.
- Time series clustering is a widely used approach for exploratory technique, and as a module in complex data mining algorithms, e.g. indexing, rule discovery, classification, and error detection [132].
- Time series cluster data can also be represented as visual images, which can help in a better quick understanding of the structure of the data, clusters, and irregularities in the datasets.

There are different challenging issues to be taken under consideration for clustering the time series data. Some of the challenges are given as [129],

- The size of the time series data can be a challenging as it will make the clustering process slow.
- The handling of high dimensional data is also a challenge as handling these data is difficult for many clustering algorithms.
- The similarity measures for clustering are also a challenging task. A similarity measure is process required to find the similarity between two time series data. This process of similarity matching for clustering is also known as "Whole sequence matching".

There are various algorithms developed to cluster different types of time series data. For using the existing algorithms (static data) for clustering of time series data, different approaches are tried that either modify the existing algorithms for clustering so that time series data can be handled, or time series data can be converted to such form that the existing algorithms can be used for clustering [128]. There are three different approaches discussed in [128] for clustering a time series data that can also be seen in Figure 2.14.

- **Raw data based clustering**: In this method of clustering the raw time series data is used for clustering, thus called raw data based approach. The major modification in this method is using the appropriate distance or similarity measure for time series data that can be used by the existing clustering algorithms for static data.
- **Feature based clustering**: This approach works by first converting the raw time series data either into a lower dimension feature vector or a various model parameters. Subsequently, applying the extracted features to a conventional clustering algorithm for clustering, therefore, this method is also called feature based approach.
- **Model based clustering**: The model-based approach uses residuals of the trained model on the time series data for clustering. For modeling the discretized form (e.g. symbolic aggregate approximation (SAX) [133]) of the data can be used, as well as the raw time series data.

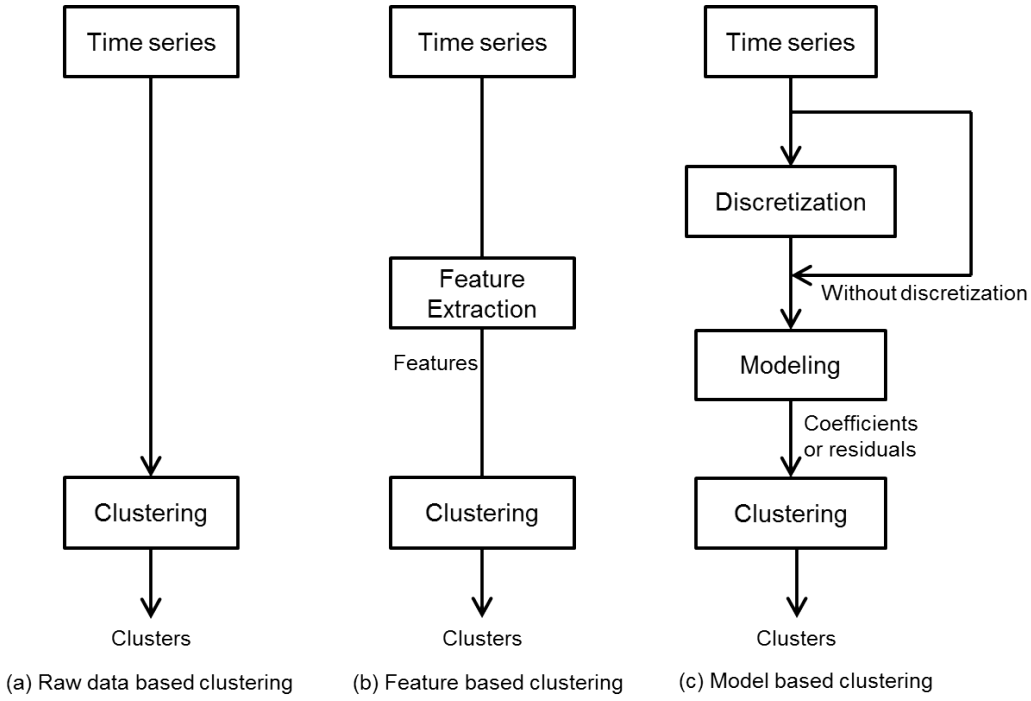


Figure 2.14: Differnt clustering techniques for time series data [128]

One of the major research topics in time series clustering is method for finding the similarity between two time series data. There are various methods proposed for time series representations and distance measures [134, 135, 136, 15, 13, 12]. In literature, the similarity matching has been mainly divided in to two main categories that are following [15],

- **Shape based similarity**: In this method, the similarity of two time series datasets is determined by comparing their local patterns
- **Structure based similarity**: In this type, the similarity of the two time series is determined by comparing their global structures.

Most of the existing approaches largely focus on finding shape based similarity for clustering, although some of the existing techniques work fine for shorter time series data, but mostly fail to produce reasonable results for the longer sequence of data sets [15]. Therefore it is required to find structural based methods for finding similarity between time series data as it will add semantics in matching of the data.

Let us consider two time series data A and B (one data is represented with red line and other with blue line) in Figure 2.15. The shape based similarity can be used to determine how much the two datasets are similar based on the local comparisons. Euclidean distance is one of the most well-known distance measure in data mining literature, and in Figure 2.15 it is used for sequence matching. The sequences are aligned in such a way that that each point one sequence is matched with same kind of point in the second sequence, i.e. the i^{th} point in sequence A is matched with the i^{th} point in sequence B. In general the Euclidean distance works well, but does not produce accurate results in a case when there is even a slight shift in the data along the time axis [15].

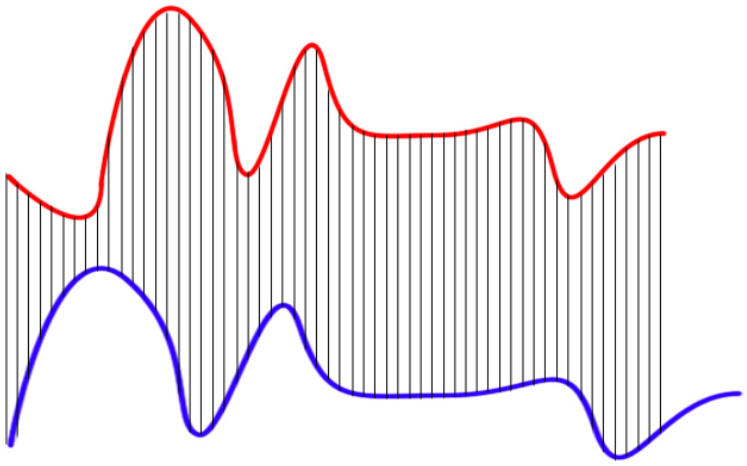


Figure 2.15: Eucleadian distance comparing two time series data

There is another distance measure method called Dynamic Time Warping (DTW) [135, 137], that overcomes the limitations of Euclidean distance method by using the dynamic programming technique for determining the best alignment to produce the optimal distance. It will need the parameter warping length for determining the best alignment. The decision of warping length select is a crucial decision as in case of selecting a large warping window, the search to become excessively expensive, and it may also cause meaningless matching between points that are far apart. Similarly, selecting a small window might fail in finding the best solution. As can be seen in Figure 2.16, the Euclidean distance can be taken as a special case of DTW, with no warping allowed. Whereas Figure 2.16 shows the time series sequences with DTW. It can be observed from the Figure 2.15 that the dips and peaks are not aligned with their corresponding points from the other sequence. The DTW is a more robust distance measure as compared to Euclidean distance, but its more computationally intensive.

One of the most appropriate methods for determining the similarity between long sequences is to measure similarity depending on higher level structures of the data. There are various structure based or model based similarities techniques proposed in literature, that can be used to extract global features e.g. trend, skewness, auto-correlation, and model parameters [138, 139]. However, it is not important to determine the relevant features, or computation of the distances having

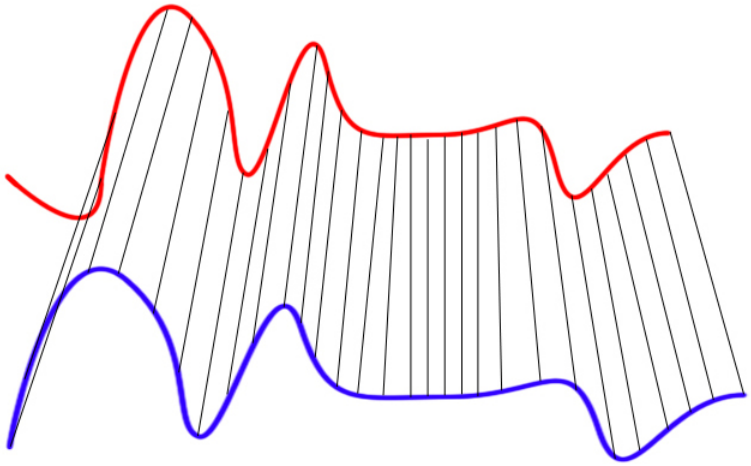


Figure 2.16: Comparing two time series data with DTW

these features, and most of the time these global features are not plenty to capture the required information for comparison of the data [15].

The representation for time series data for finding similarity is Discrete Fourier transform (DFT) [140]. DFT finds the approximation of the signal by using the linear combination of basis functions. The coefficients of DFT represent the global contribution of the signal. Other than this, there is another well-known method for representation of time series data called Discrete Wavelet Transform (DWT) [134]. The Wavelets uses mathematical functions for representing the time series data, or, some other functions showing relationship of the averages and differences of a prototype function, also known as mother wavelet. The authors in [136] have shown that DFT and DWT presented same performance in terms of accuracy.

There are methods where learning auto-regressive moving average (ARIMA) model on the time series have been proposed, while using coefficients of the model coefficients as the feature for representing time series data [141]. A deformable Markov Model template can also be used for temporal pattern matching, where the data is transformed to a piece-wise linear model, but requires many parameters [142]. The statistical features of time series data can also be used for time series data representation e.g. mean, variance, skewness, and kurtosis. Subsequently these statistical data can be used for classifying the data using multi-layer perceptron (MLP) neural network [138].

The Symbolic Aggregate Approximation (SAX) is a type of Piece-wise Aggregate Approximation (PAA) representation and can be used to improve the speed and usability of several analysis techniques [133]. The Symbolic aggregate approximation (SAX) representation can also be used for finding the similarity between the time series data having same behavior [133]. Furthermore, another method was proposed by authors in [133] called as bag of words technique. The bag of words representation was used with hierarchical clustering for improving the algorithm for finding the similarities between the different time series data. Additionally, in [133], the comparison of several algorithms (Euclidean, DTW, wavelets) with proposed method called bag of words method using hierarchical cluster for finding the similarity between different time series is given. The authors have concluded that the bag of words approach performed better for finding the similarities in the time series data as compared to the other methods.

2.5 Cluster Validity

It is required to check the quality of the clusters that how different elements are distributed in clusters and validate that the clustering has meaningful results, after clustering. There is extensively use of cluster analysis techniques for exploratory data analysis, with applications in various fields such as speech processing [143], information retrieval [144], and Web applications [145]. Clustering has been used widely used as a basic tool with modification for different application [146]. As clustering is unsupervised method, which means the number of clusters is an unknown parameter, and the data analyst has very less knowledge of the data. Therefore, it is necessary to evaluate different clustering algorithms, and find the optimum number of clusters for cluster analysis. As clustering is defined as grouping data points in such a manner that same elements are out together and the dissimilar elements are put in other clusters [146]. Subsequently, this rule makes the sketch for the techniques of cluster evaluation. Some of the existing clustering algorithm partition data into a predefined number of clusters such as k-means clustering algorithm.

The use of cluster validity can help in determining the accurate number of clusters in a data set. It is very difficult to define a good clustering criterion as it requires detail understanding of the data. For this purpose clustering can be used as one of the principal tools for understanding the data [147]. There are various cluster validity indexes techniques proposed [146]. Generally, the cluster validity techniques are classified into two types' i.e.,

- **Internal indexes**: The internal indexes are usually based on the basic information about the data
- **external indexes**: The external indexes are based on the prior knowledge about the data.

One of the solutions used for determining the number of clusters is by finding a knee point among the validity index values with different number of clusters. The knee point or also called as elbow is the number of clusters having sharp change at the index values. The validity indexes representing either the minimum or maximum value are preferred. Though, the possibility is always there for the validity index to have several local minimum or maximum points.

Information criteria can be used for determining the number of clusters in a dataset, such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), or the Deviance information criterion (DIC) [148]. Other than this cross-validation process can also be used to analyze the number of clusters in the data. The data is partitioned into parts in this technique. After partitioning, each part is used as a test set on in its turn. The remaining parts (except the test part) are used for clustering model as training sets, and the value of the goal function calculated for the test set. After calculation of these values, they are averaged for each alternative number of clusters. The cluster number that minimizes the test set errors is selected as the optimum number of clusters in the data [149].

The decision of the optimal number of clusters is an important issue generally in unsupervised method and specifically in case of hierarchical clustering, as the clustering algorithm can give best results where the difference in inter-cluster elements can be as less as possible, whereas the difference between the intra-cluster can be as greater as possible [150]. One of the common graphical methods that can be used for cluster evaluation uses the plot an error measurement against the numbers of clusters. In this method the position where the plot creates "elbow" in the graph can be taken as the number of clusters, as usually the "elbow" occurs at the sudden

decrease in the error measurement [151]. The Figure 2.17 shows the example taken from [150] with two graphs, where the Figure 2.17(a) is representing the data, and the Figure 2.17(b) is showing the number of clusters with error measurement. It can seen clearly from the Figure 2.17(b) where there is sudden change(elbow) and can be considered as important point for finding number of clusters. Therefore, the optimum number of two clusters will be selected, which is also clear from Figure 2.17(a)

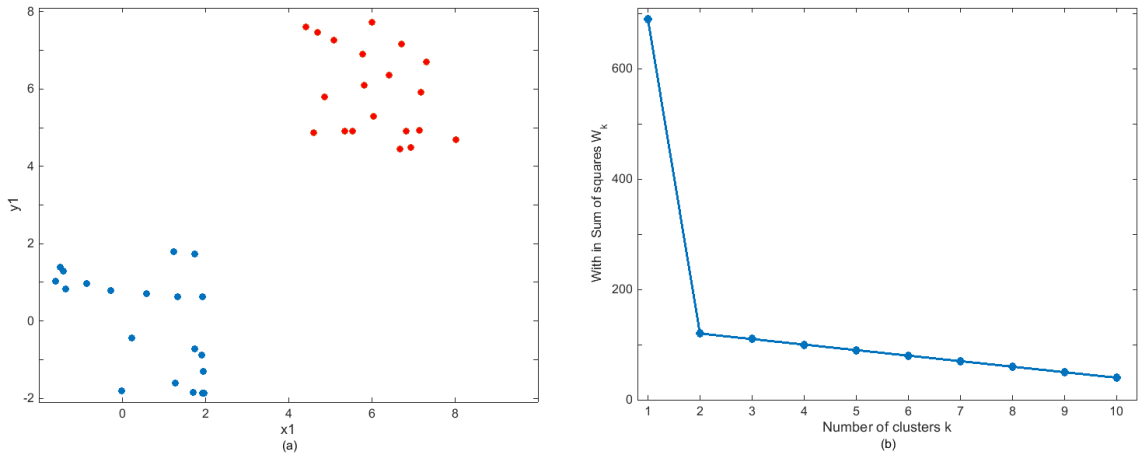


Figure 2.17: Example of finding optimal number of clusters with elbow method [150]

There are other methods that can be used to find optimal number of clusters in the data such Silhouettes criterion method [152], Davies-Bouldin criterion method [153] and Calinski-Harabasz criterion method [154]. Other than these techniques, there is a method that is found in the literature is based on gap statistics analysis where the gap criterion find the optimal number of clusters by estimating the "elbow" location as the number of clusters against the largest gap value [150]. The gap value can be defined as [150]

$$Gap_n(k) = E_n\{log(W_k)\} - log(W_k), \tag{2.4}$$

here E_n is the expected value, n is the size of sample, k is the number of clusters that is being evaluated, and W_k is the dispersion measurement within the cluster and can be find as

$$W_K = \sum_{r=1}^{k} \frac{1}{2n_r} D_r, \tag{2.5}$$

here n_r represents the count of data points in the cluster r, and D_r denotes the sum of the pairwise distances for all data points in the cluster r.

The Gap distance has been calculated for the same data as shown in Figure 2.17(a). The gap distance along with the number of cluster have been drawn as can be seen in Figure 2.18. The x-axis in Figure 2.18 represents the number of clusters (k), whereas, the y-axis display the calculated gap distance. It can be observed in the Figure 2.18 that the gap distance is maximum at cluster number 2 at the a-axis. Therefore, It is a clear indicator that the optimum number of clusters is two in the given data. So, the optimum number of cluster for the given data set is two, that means that the intra-cluster elements will have minimum possible distance, and the inter-cluster elements of the cluster will have the maximum possible distance with two clusters.

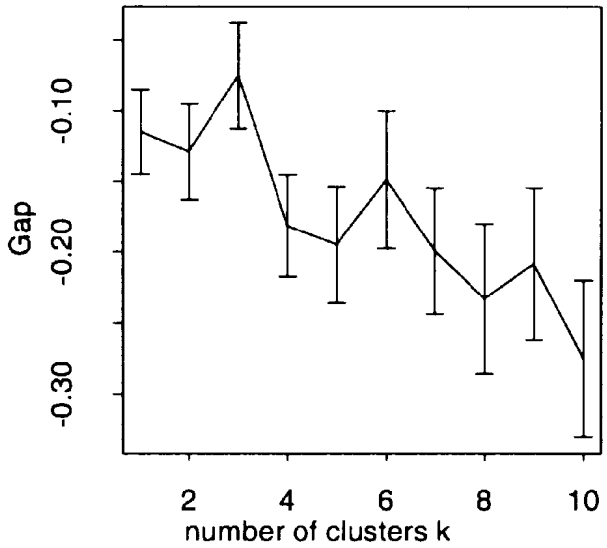


Figure 2.18: Example of finding optimal number of clusters with Gap method [150]

3 Monitoring Framework and Overview of the Methodology

This chapter gives an overview of the monitoring framework adopted, and the methodology that has been used in this research. The first section explains the monitoring framework for the different energy systems that has been under observation, and data from these monitoring framework has been used for validation of the proposed methodology. In addition, the description of the selected parameters that are used during the research and its position in the monitoring system is discussed. These parameters are used for evaluation of the systems, therefore selection of these parameters is a critical decision to make. The operational data of chillers (adsorption and absorption), and ventilation system has been used for validating the proposed methodology. In the last section an overview of the methodology used for faults detection and diagnosis in energy systems of the building, has been discussed which has been adapted for the research in this thesis.

The main focus of this research is to find a method that can automatically detect and diagnose faults in the operational data of energy systems of the building. The data analytics and machine learning is one of the prominent research area that can help in automatically detecting the different patterns in buildings data. Therefore, a work flow for automatic detection and diagnosis of energy systems, using data analytics techniques and machine learning algorithm has been proposed. The main contribution of this dissertation is to develop a work flow using a robust set of algorithms for automated data quality check and data analysis of the energy system with little configuration efforts. The proposed work flow method is more appropriate for the analysis of the energy systems in general. The different phases of the work flow start with data sanitation, where solution for various issues have been given such as automatic duty cycle detection, validation of data using rule of first principles and visualizations, limit checks, outlier detection, and missing data imputation. In the second phase the method for detection of different anomalies or noise added to the data, during recording of the data has been discussed. In the last phase clustering is used, in order to find the various operation patterns in energy systems of the building. After clustering the data, the cluster having faulty patterns have been targeted and additional information is included in analysis for diagnosing of the faults in the energy systems of the buildings. The overview of the methodology can be summarized as in Figure 3.1, where the rectangles represent process that are focused on and circles are the proposed algorithms for the process.

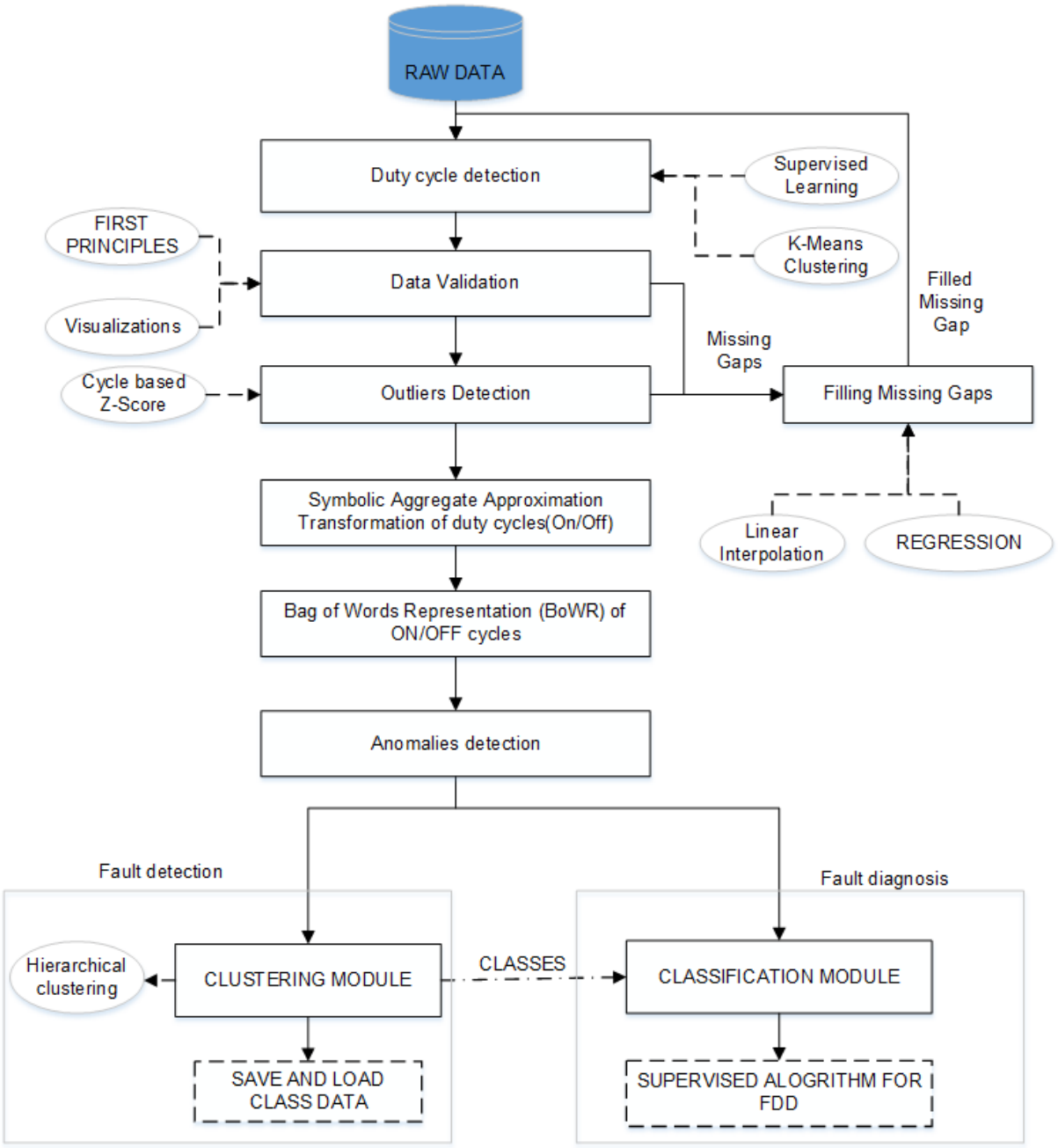


Figure 3.1: Methodology Adopted for Research

3.1 Monitoring Framework

This section discusses the different types of monitoring frame works used in this research. The data under consideration is taken from chillers (adsorption/absorption), which is manufactured by company Pink chillers, and ventilation system of the building known as Energybase [155]. The data is fetched from the energy systems and is then stored in a monitoring and analysis system OpenJEVis [25]. OpenJEVis is an open source solution providing functionality for monitoring and analyzing the energy data. It provides data storage feature and different visualizations of

data. Additionally it also provides a meta-data structure for tagging different states of data like faulty etc. The benefit of Jevis is that the data is recorded directly into the storage, the a pool of FDD algorithms can be applied and relevant meta information can be saved in it, and the relevant information can be viewed by user like facility manager without the knowledge of the whole process. The scope of this thesis is limited to Fault detection and diagnosis (FDD) module.

3.1.1 Monitoring System and Data Parameters

This section explains the monitoring system and the parameters that are used for the evaluation of the energy systems of the building. There are two types of energy systems used for the experimentation and testing of the proposed methodology. The energy systems include chillers (Absorption and Adsorption), and air handling unit.

The chillers used are solar driven cooling system. Therefore the hot water is produced by using solar energy for generating cold water, which is again used for cooling for cooling the space in the building. The cooling produced are used for maintaining comfortable environment in building facilities like meeting rooms etc. Furthermore, the cooling capacity is also used for cooling facilities of food products. The design of the system is shown in Figure 3.2, showing the three temperature

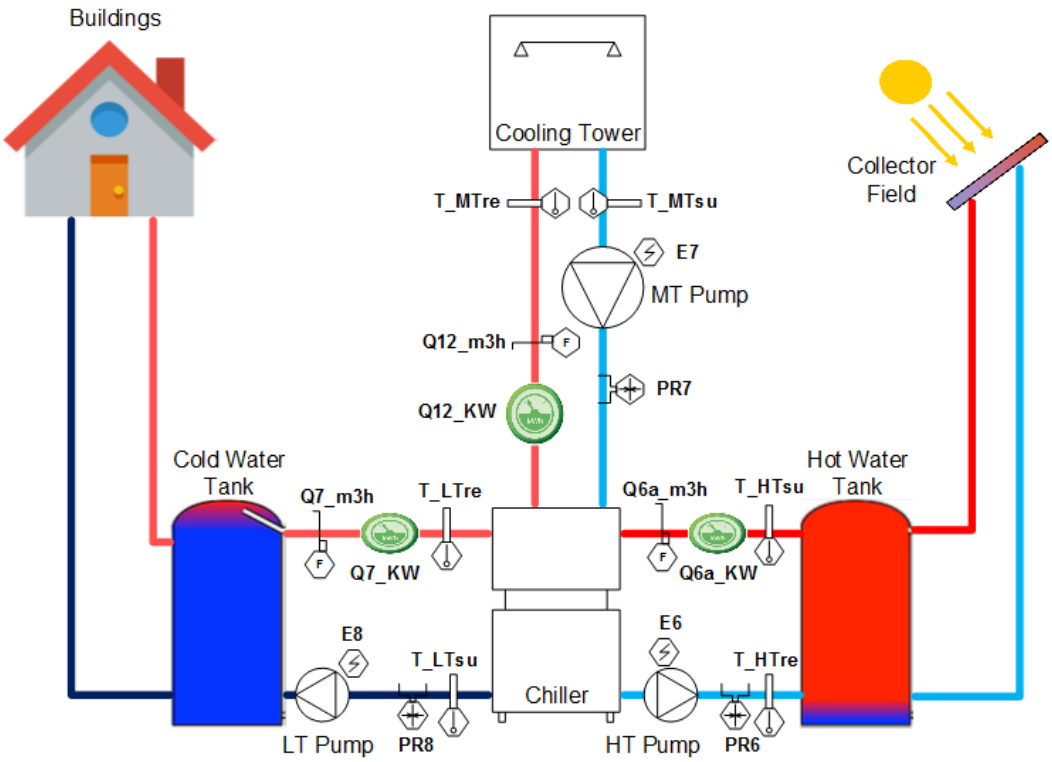


Figure 3.2: System Design

cycles i.e. Low Temperature (LT), Medium Temperature (MT) and High Temperature (HT) cycles along with the installed sensors that are used for analysis. The HT cycle is representing all the required sensors that were required for the evaluation of the hot water cycle performance. Similarly the LT cycle is showing the required sensors for the evaluation of cooling produced by the chiller, and MT cycle display the sensor responsible for taking the undesired heat produced

during the process to the environment. Figure 3.2 show the the different electrical consumption, pressure, temperature, energy meter and flow reading meters for the three cycles which are the parameters taken under consideration for this research.

For the scope of the thesis a total of 18 parameters were selected for evaluation of the chillers, after a detailed discussions with the experts in the field. The description of each parameter is given in Table 3.1. The naming convention of IEA SHC Task 38 has been adopted for the considered parameters and to provide a unified monitoring procedure. The unified monitoring procedure will help in estimating the primary energy savings of monitored SHC systems and will enable to do the comparison between different monitored SHC systems.

Table 3.1: Parameters Description

Sensors	**Description**
$E6$	High Temperature (HT) electricity consumption meter.
$E7$	Medium Temperature (MT) electricity consumption meter.
$E8$	Low Temperature (LT) electricity consumption meter.
$Q6a\_m3h$	HT cycle Flow (water) reading
$Q12\_m3h$	MT cycle Flow (water) reading
$Q7\_m3h$	LT cycle Flow (water) reading
T_HTre	HT cycle temperature on return side.
T_HTsu	HT cycle temperature on supply side.
T_MTre	MT cycle temperature on return side.
T_MTsu	MT cycle temperature on supply side.
T_LTre	LT cycle temperature on return side.
T_LTsu	LT cycle temperature on supply side.
$Q6a\_KW$	HT cycle Energy consumption reading
$Q12\_KW$	MT cycle Energy consumption reading
$Q7\_KW$	LT cycle Energy consumption reading
$PR6$	Pressure in HT cycle
$PR7$	Pressure in MT cycle
$PR8$	Pressure in LT cycle

The data from air handling unit of building is also taken under consideration during the testing and experimentation process. The ventilation system supplies the fresh air to the lecture rooms, office spaces and common area of the building called as Energy Base [155]. The two facilities i.e. Cooling and heating are provided in a limited possible way (2 Kelvin)over or under the room temperature, depends on the heating or cooling mode. The main components of the ventilation system are

- Heating Coil.
- Cooling Coil.
- Heat Recovery system.
- Supply and exhaust Fan (with frequency inverter for controlling the speed of fans).

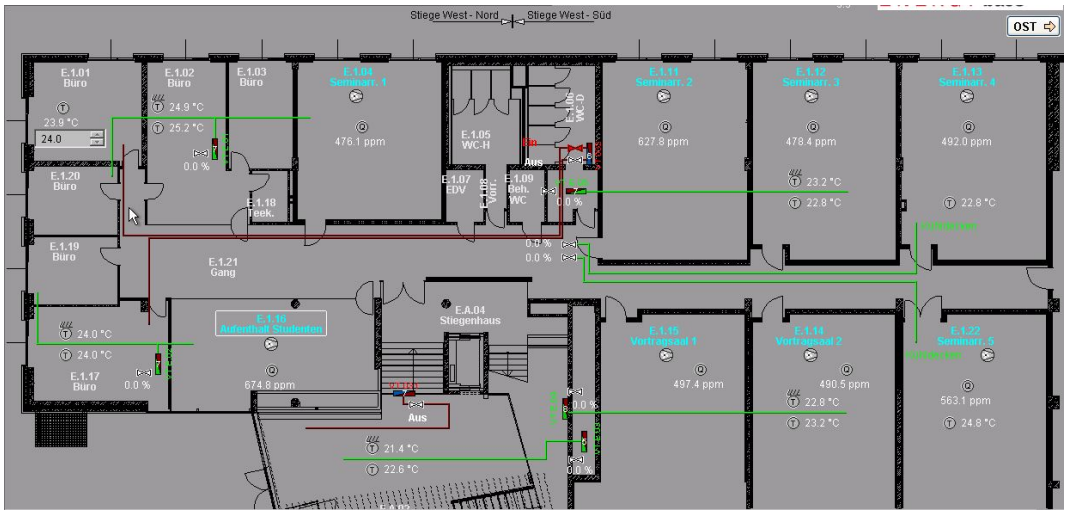


Figure 3.3: Floor plan of the building for the ventilation system

Figure 3.3 shows the floor plan of the ventilation systems. The floor shows the different rooms and placements of sensors that are used for controlling the working of ventilation system.

In order to manage the ventilation system, two different control strategies of the volume flow rate for ventilation system are used. At the first level, the ventilation system is controlled with flow pressure, using the speed functionality of the fans, and at the second level, in rooms with carbon dioxide sensors using variable air volume (VAV) box. Figure 3.4 shows the supply and exhaust fan placement and the placement of sensors recording the pressure and temperature in the supply and exhaust duct. The Figure 3.4 shows the how the air is taken in, and send out of the building.

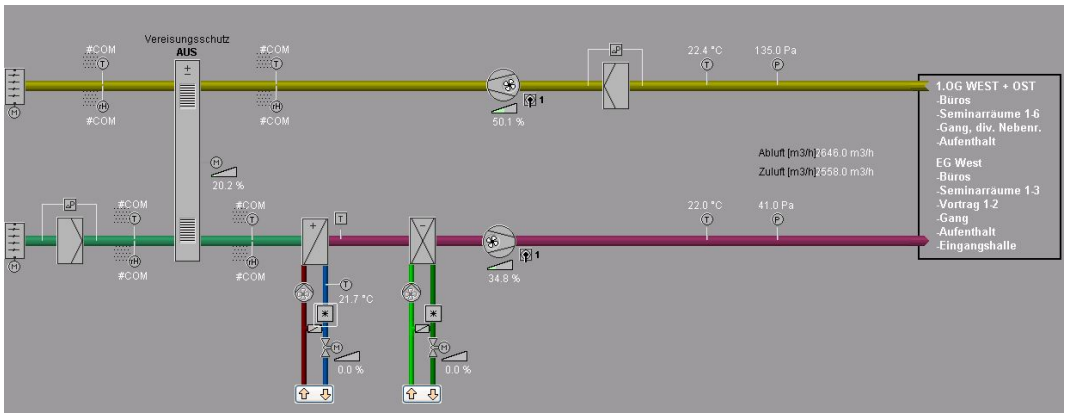


Figure 3.4: Air handling unit showing the exhaust and supply fan along with monitoring sensors

Figure 3.5 shows the variable air volume boxes that control in flow and out flow of air in the rooms of the building. The VAV boxes are controlled with the air quality of the room such as carbon dioxide (CO2) sensors, that are fixed in the rooms. The VAV values are represented with percentage (%), showing the amount of the air that the VAV has allowed in or out of the room. The VAV are fixed on both ducts for controlling the fresh air that has been brought in, and the existing air in the building that has been send out of the room.

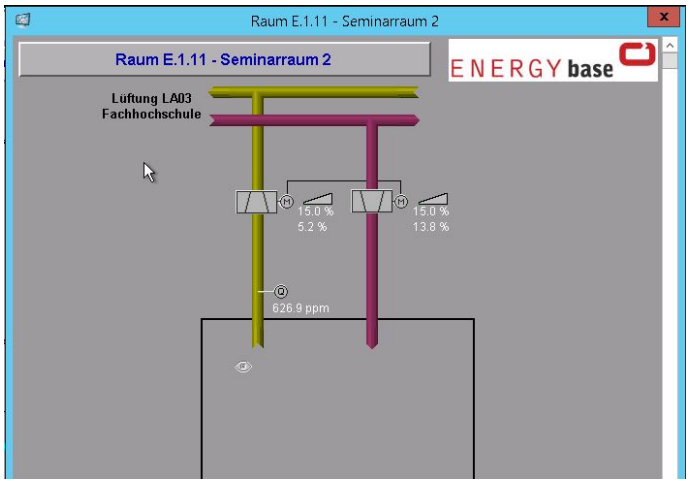


Figure 3.5: Variable air volume flow (VAV)

The air handling unit has various parameters inside room, consisting of different types i.e. temperature (Temp) sensors, relative humidity sensors (rH) and carbon dioxide (CO2) sensors, supply and exhaust pressures, supply and exhaust volume flow, and variable air volume signals etc. The sensors are place in two rooms of building, represented with E012 and E015 and inside placement in room are denoted with left (L), right (R), front left (VL), back right (RH), whereas the sensor placed in the center of room have one additional character added to their name showing its position as up (O), middle (M), and down (U). The monitoring policy of the air handling unit is given in the Figure 3.6.

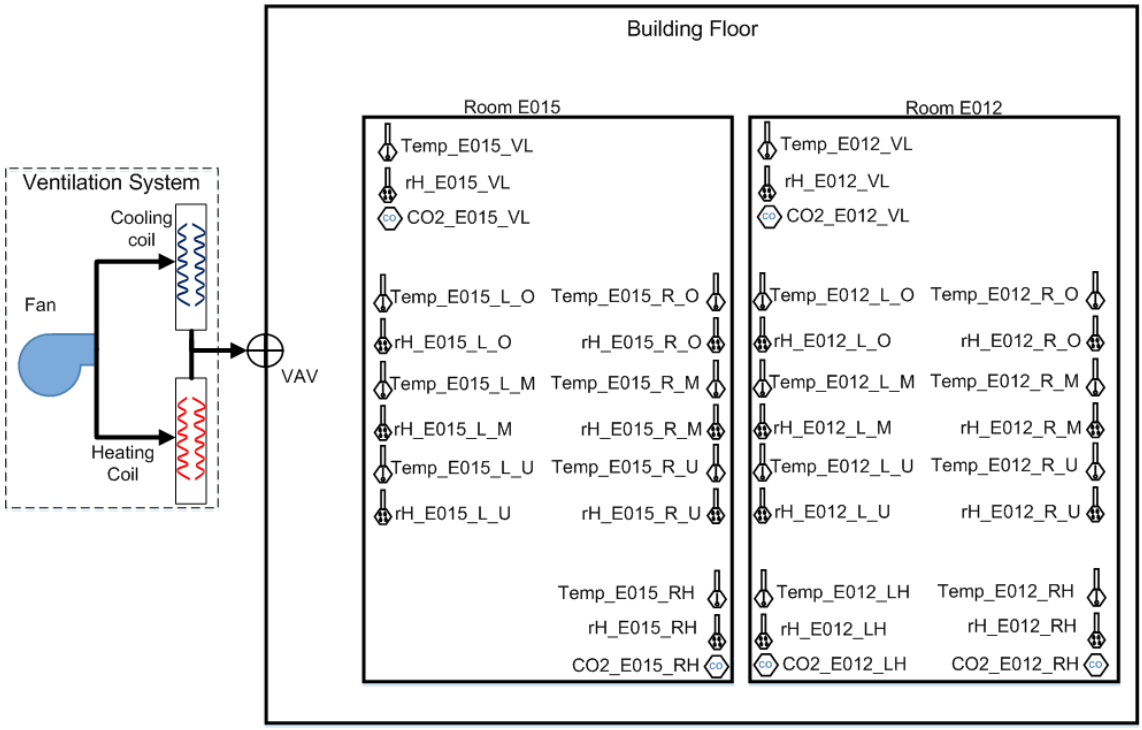


Figure 3.6: Air handling unit monitoring inside a room with temperature, relative humidity and carbon dioxide sensors

3.1.2 Implementation Tools

The concepts suggested in this thesis are implemented using the KNIME tool along with the MATLAB. KNIME stands for "the Konstanz Information Miner" [156]. It is an open source data analytics platform, having additional features that include reporting and integration platform using several other tools. The KNIME uses the modular pipelining concept to integrate various components from different other tools for machine learning (such as WEKA, R) and data analytics at the same place. The KNIME has graphical user interface that allows creating nodes for various tasks such as data preprocessing, modeling, data analysis and various types of visualizations. Figure 3.7 shows the graphical user interface of KNIME.

The main features of the KNIME can be summarized as following

- The KNIME allows its users to use the graphical user interface for generating new data flows. It also allows executing either some part or all analysis steps at once. The results can be viewed later using different built in visualization. It has the features of having interactive views.
- KNIME is implemented in Java and is based on Eclipse, thus it allows using its extension mechanism for adding other plugins for providing additional functionality such as open source tools for machine learning algorithms from Weka, and the statistics package R project. In addition it also provides nodes for running code in Java, Python, Perl etc. Other than this it provides additional plugins for the fields relevant to the Text mining, Image mining, and time series analysis as well.

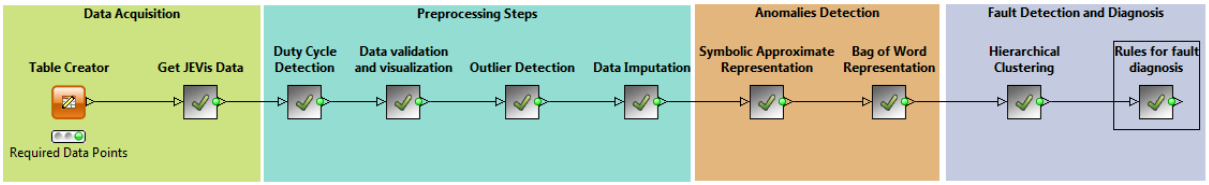


Figure 3.7: KNIME workflow

3.2 Overview of the Methodology Used for Detection and Diagnosis in the Energy Systems of the Building

This section gives an overview of the methodology adopted for this research. There were various aspect taken under consideration while analyzing the data for detailed analysis of the energy systems in buildings for the purpose of finding a work flow for automatic detection and diagnosis of the faults. The main idea covered in this thesis can be categorized in three concepts, i.e. data sanitation, fault detection, and diagnosing the detected faults as can be seen in Figure 3.8.

Nowadays energy systems installations (e.g. solar cooling) devices have the problem, that the employed monitoring system along with existing methods for fault detection is not satisfactory. They are able to accurately detect few faults and are not able to detect and diagnose many fault scenarios correctly yet. One of the reasons might be the faulty sensors that records and transport wrong values. These wrong measurements will result the devices to shut down, although the

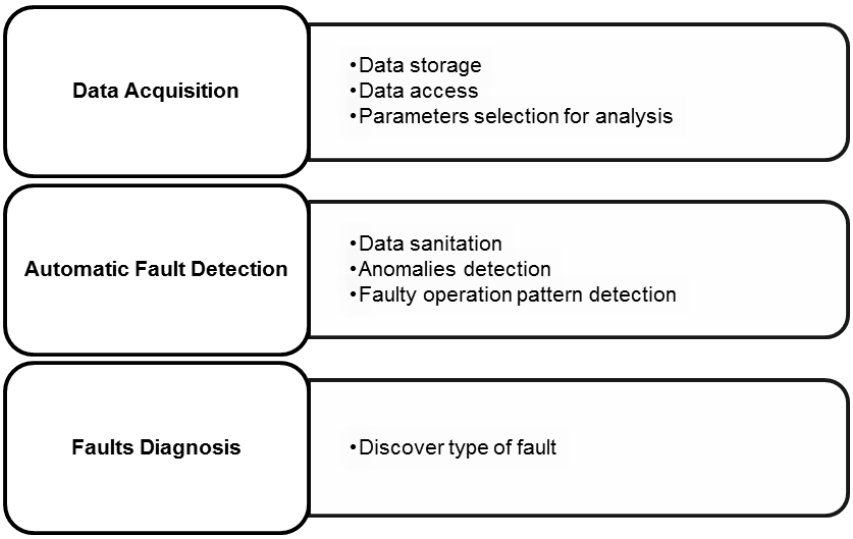


Figure 3.8: Main sections of the methodology

system has no problem while in operation. On the other side, there can be situations where sensors report the measurements correctly, but the monitoring system is not able to detect these faults. In most such situations only the trained engineers can detect faults after a time consuming investigations of the relevant measurements. The circumstances can be troublesome specifically in case if dealing with solar based energy systems in buildings, as these systems are operational for a few hours a day and in case of fault the solar irradiation could not be used efficiently because of system faults. Therefore, it is required to detect faults and resolve it quickly for achieving good performance of the energy systems.

Furthermore, the practical experience, of handling the historical monitoring data from sensors, has revealed that the data is mostly inaccurate and incomplete [8, 9, 10]. There are different issues observed with the data such as modifications in the configuration that are not correctly tracked, improper calibration problems of sensors, outliers in the data (data points that do not represent the behavior of the system), and some other issues in the data storage structure. Therefore, data sanitation is required for validating the date before the detailed analysis of the data. The data sanitation is the process of improving the quality of the data that can be further used in analysis. It has been observed that the conventional analysis method using the calculations found in standards such as EN 15316 produces unsatisfactory results if the quality of the data is low [11]. So, a method of data sanitation is need that can sanitize data automatically using as less as possible input of the expert in the field.

The sensors are generally deployed in the fields to monitor and record the real life phenomenon such as temperatures, electricity consumption, in the energy systems of the buildings. Usually, these sensors in the field are operating in hostile environments; thus the sensor recorded data has high probability of anomalies existence in the data. The detection of such anomalies is necessary and is a challenging task due to the large amount of data recorded in the energy systems of the building. Furthermore, in order to evaluate the performance of various energy systems, it is required to have an automatic method for fault detection in the operational behavior of the energy systems. The behavior of the buildings energy system is quite complex to capture the faults pattern, therefore a method is needed for automatically detection of faults in the operation of systems without the help of experts. One of the easiest approaches, adopted by the experts

in the field, is to use different visualization tools. But, the huge amount of data recorded during monitoring, makes it difficult for the experts to have a detailed performance analysis of the buildings manually or with simple visualizations. Moreover, there is a high chance for overseeing some important patterns in the data that may lead to faults in the different components of building, causing reduction in the energy efficiency. The methods that can automatically extract the different patterns in buildings data will help the experts to get more insight of the different parameters of the buildings energy usage and other processes such as factors causing faults in different components of the buildings.

The diagnosis of the faults in also needed as it will help to know the reason for the low performance of any energy system in the buildings. The correct diagnosis will help to remove the fault, thus will help in the reduction of electricity consumption of the energy systems in the buildings. The sections of the methodology are summarized as following:

3.2.1 Duty Cycle Detection

The duty cycle detection has been done using supervised learning methods such as k-nearest neighbors (KNN), support vector machines (SVM) and random forest trees. The main issue with using the supervised learning is the availability of data that is labeled for generating the models.

In order to have a robust method for automatic detection of the duty(operational) cycle, the behavior of the different parameters such as temperatures, flows, energy readings and pressures of various energy system are analyzed. The supervised It is observed that the data has two different patterns showing the operational (On) and non-operational (Off) state of the energy systems. Therefore, on the basis of this assumption that the energy system behavior pattern will different in the two states, a method has been proposed using k-means clustering algorithm that clusters the data in two clusters representing Operational (On) and non-operational (Off) cycle.

Furthermore, k-means is also used to find the occupants presence in the building using the ventilation system information with the air quality parameters such as humidity, temperature, carbon dioxide level inside the room. As the air quality will differentiate with occupants presence inside the building, therefore, k-means with two clusters was able to detect the occupants presence inside the building.

3.2.2 Data Validation

After the automatic detection of On/Off cycle, the data is validated using first principles and some visualizations such as histogram visualization are suggested for limit check are proposed. The validation of data is required to make sure that the data used for detailed analysis should exhibit the required behavior in order to save time and energy in the next steps. Otherwise the next steps will not benefit as "Garbage in, garbage out". Figure 3.9 was amended and used to see the data availability at different times of the year. The calendar view does not need any domain knowledge and provides a good overview of the data.

It can be a good starting point for further detailed analysis and investigations of the faulty behavior, since the problematic time periods can be identified easily. In Figure 3.9 the data of chiller is shown with that represent the operation state of machine, during the cooling season in summer. One of the example for further investigation can be the data gaps such as on 6 September 2010, can be conveniently visualized.

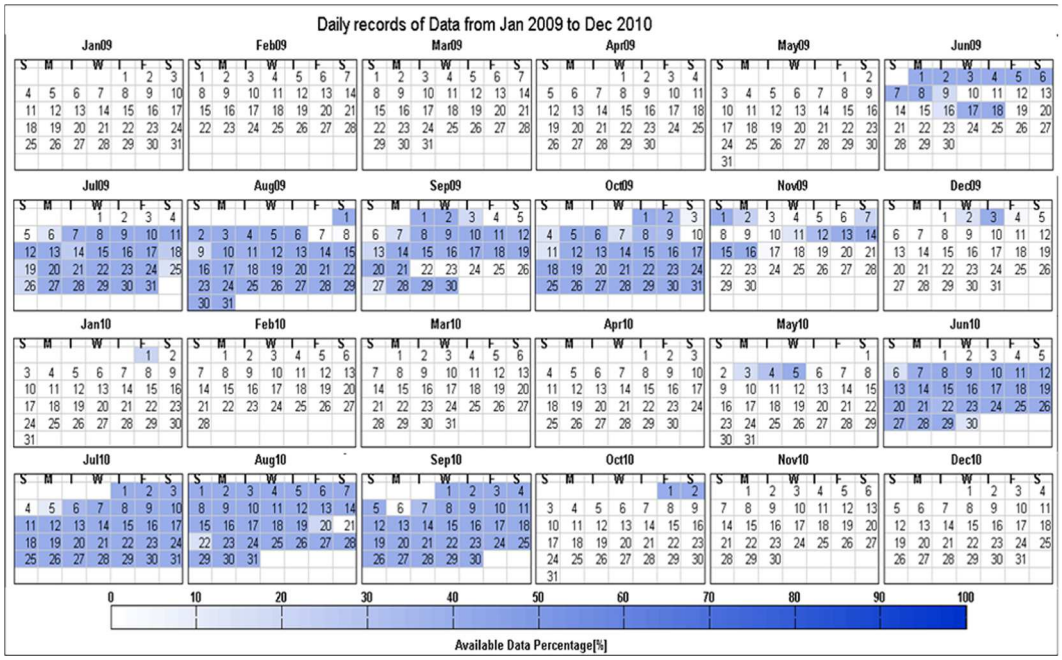


Figure 3.9: Calender view of the availability of the data

3.2.3 Outliers Detection Using Duty Cycle Based Z-Score with Expectation Maximization Clustering Algorithm

As the behavior pattern of the energy system varies in two states i.e.operational (On) and non-operational state (Off), therefore this information is used for automatic detection of outliers. The outliers can be defined as those data point that do not correspond to the expected normal behavior of the energy system . So a method for automatic outlier's detection is proposed using z-score normalization based on the duty cycle (On/Off) state of the energy system, and subsequently, clustering the data by expectation maximization clustering algorithm. The duty cycle based z-score normalization helps in highlighting the outlier's for clustering algorithm to cluster these points away from the normal behavior of the data. Thus helps in finding the invalid data points in the data. The results of proposed method were compared with other state of the art methods and it was found out that the proposed cycle based z-score method was able to detect outliers more accurately.

3.2.4 Filling Missing Gaps

This section discussed the two methods i.e linear interpolation and regression, which have been used for data imputation of the missing data. There are missing gaps in the data due to various reasons such as communication failure or sensor failure etc. The decision of replacing these missing gaps (data imputation) is very critical as this can affect the analysis of any system. The simple most approach can be to ignore these parts where the data is not available. But the decision of ignoring the data is possible only when smaller portion of data is missing. There are other methods that can used for data imputation such as using mean, median, nearest neighbor, interpolation and regression depending upon the nature of the data. As there will be some missing gaps introduced because of the data validation and outlier detection step, therefore it is required

to fill these gaps. In this research a method for data imputation has been proposed, where shorter gaps are filled with interpolation, but larger gaps are not filled.

3.2.5 Anomalies Detection

This section handles the detected anomalies in the data. There are various types of anomalies in data of sensors recording, as sensors usually are deployed in fields for monitoring. A method has been proposed to automatically detect anomalies patterns in sensor data, that is recorded in energy systems of the building. A crucial issue with monitoring data in energy systems is the data quality of the recorded data, due to problems in commissioning, transmission, storage or the sensor itself, the data is often not feasible for further analysis, thus impairing the ability to detect faults in the operation of the energy systems. The focus lies on finding a method that covers all different possibilities of error patterns and has a high degree of automation so that it requires close-to-none configuration. The two types of anomalies found in the analysis of the data are *Constant* patterns and *Noise* patterns in the data. The data is subjected to k-means clustering for the unsupervised classification of On/Off cycles.

After the detection of duty cycle, the data is normalized in next phase using z-score, followed by discretization process using symbolic aggregate approximation. Each cycle data is first broken down into M non overlapping sub-sequences, in a uniform manner, just like the example illustrated in Figure 3.10, wherein the partitions are represented by alphabets a, b, c and d. The SAX representation of the example cycle given in Figure 3.10 will be (aaaaaabbbbbbbbbbbbc-ccccccccccccccccccbbb). This process is called as chunking, and the period (x-axis) can be of

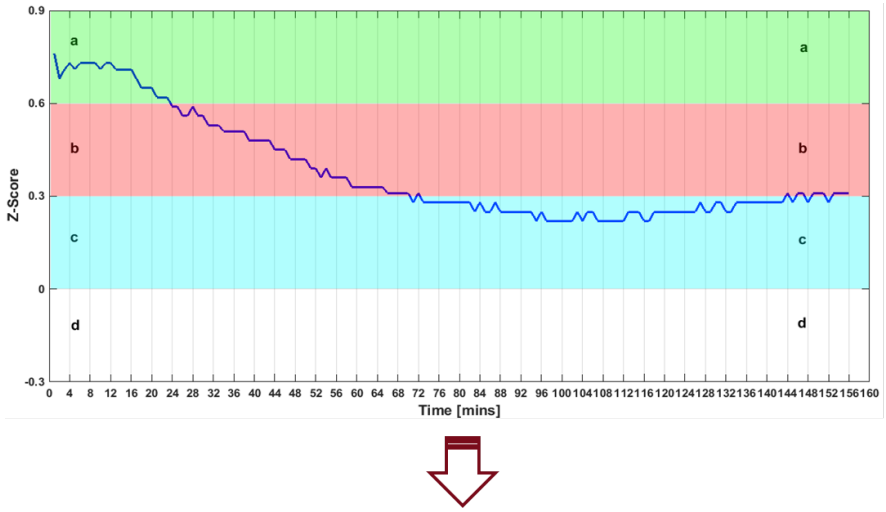


Figure 3.10: SAX Transformation example with M=4 and P=4

different time length (P) depending on the application where it is used [12]. The value of P is taken as 5 minutes in this research. The symbol of each data point is assigned according the breakpoints. The number of break points (M) taken for this research is 60.This transforms the data for each cycle to symbols. The SAX representation is specific for each a length of each cycle. In order to generalize the symbolic representation for each cycle with different lengths, the

BoWR is used. The process of bag of word representation conversion can be seen in Figure 3.11. After discretization of the duty cycles (On/Off), a bag of word representation (BoWR)of each duty cycle (On/Off) is created. In order to handle the different time length cycles a better idea is to normalize, i.e. use relative frequencies.

For finding the anomalies in the sensor data, the bag of word representation pattern of consecutive duty cycles are compared. As it is expected that energy systems in building perform differently in operational and non-operational cycles, therefore in case of having same patterns in consecutive cycle is an indicator of anomaly in sensors data.

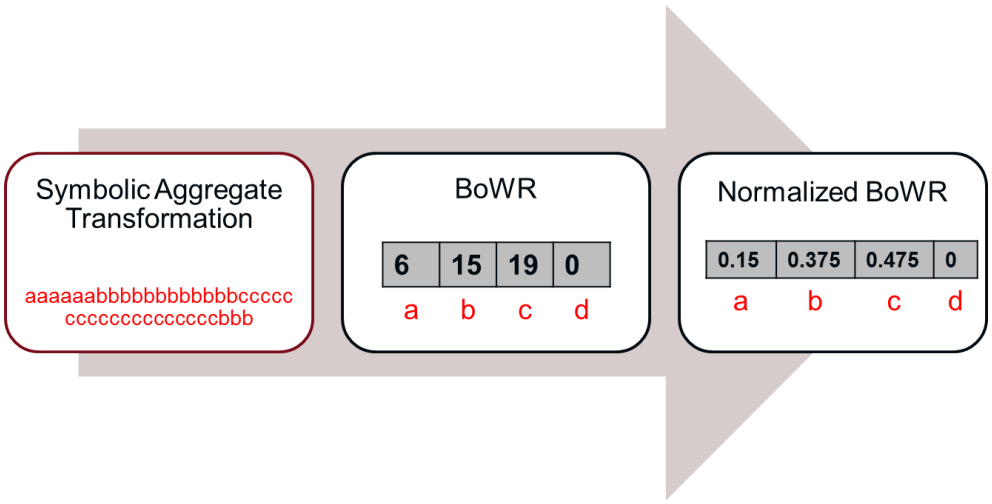


Figure 3.11: BoWR transformation example

3.2.6 Finding Fault Patterns During Operation in Data

This section discusses the various different patterns of faults in energy system of the buildings. There were different types of energy systems used for analysis i.e. air handling unit and solar cooling systems.

3.2.6.1 Air handling Unit Analysis

In order to analyze the operation of the air handling unit various types of visualizations are used. The k-means clustering is used to find the occupants presence in the building. Various visualizations were used to check the operation of the ventilation system in the building. During the analysis of the Ventilation system, it was found out that the ventilation system was running even when there were no occupants in the building. Therefore, energy consumption of the ventilation system can be reduced, if the ventilation system control is managed automatically by using the occupants presence in the building information.

3.2.6.2 Solar Cooling Systems (Chillers)

This section elaborates the method used for finding the various patterns in the data of On cycles of the solar cooling systems. In order to automatically detect faults in the operation of the energy

systems and to find various patterns in the energy system operation, a bag of words representation (BoWR) with subsequent hierarchical clustering has been proposed. The method uses the duty cycle detection information by the k-means clustering algorithm. The operational (On) cycles are of greater importance for finding the performance of energy systems and further in detecting and diagnosing faults. Additional features are suggested for achieving better results of clustering algorithm. The hierarchical clustering uses the BoWR of the On cycles of each feature for finding the various operational patterns of the energy system. The hierarchical clustering technique groups the data over a different scales by creating a cluster tree which is also called dendrogram. The dendrogram shows a multilevel hierarchy of clusters, where the clusters (groups) at one level are joined together as clusters at the next level. This property of hierarchical clustering allows to decide the level of clustering that is most appropriate for the task it is used. In order to decide the best level or optimum number of cluster, the gap statistics have been suggested. The cluster having the duty cycles with less average performance and longer duration is considered to be possible area of faults.

Figure 3.12 shows the steps involved in finding the different patterns in the data of energy sytems in buildings. It can be summarized as follows,

- Find the On/Off cycles of adsorption chiller using k-Means clustering algorithm as discussed in section 4.1.2.2 of the thesis.
- Create a bag of word representation (BoWR)for all the On cycles. The BoWR will contain the selected features patterns.
- Use hierarchical clustering to group the similar cycles having same patterns.

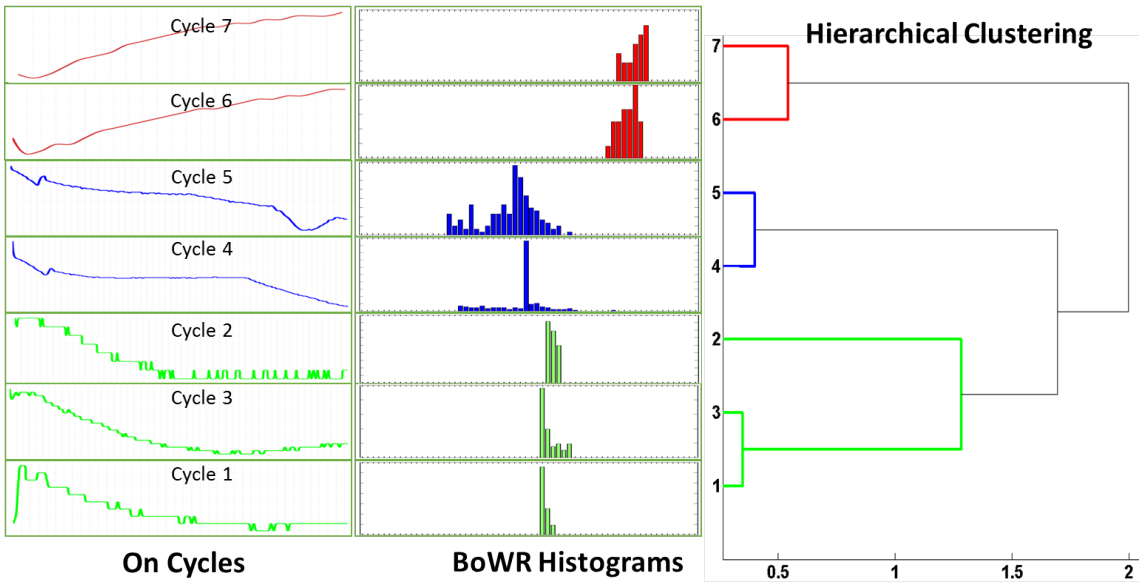


Figure 3.12: Fault and detection and diagnosis using hierarchical clustering

3.2.7 Diagnosis of Faults

The portion explains the method adopted for diagnosing the fault patterns. After clustering the data, additional information is added to each cluster such as average coefficient of performance

(CoP) of the cluster and average time of On duty cycles in a cluster. Furthermore, additional parameters of operation have been integrated into the analysis, enabling to diagnose these faults after comprehensive discussion with experts in the field, the following additional parameters of operation had been integrated into the analysis:

- Evaporation Pressure set point of absorbent($P_{Eva-setpoint}$).
- Low pressure of the system(P_{system}).
- Injection Value status.

An algorithm using various rules have been suggested for the diagnosis of faults in the operation of the energy systems. The algorithm had targeted five different types of faults of water based (adsorption) cooling chiller systems i.e.,

- Low temperature (LT) pump not proper functioning.
- Medium temperature (MT) pump not proper functioning.
- Cooling tower not performing functionality.
- Injection valve fault.
- Solution pump fault.

4 Data Sanitation

This chapter discusses the various steps of the work-flow for data sanitation and analysis of operation data. The proposed work-flow steps are performed ahead of the detailed analysis of energy systems in building data. The process of improving the monitoring data, in order to increase the data quality is called as data sanitation. The data sanitation work-flow will help to add the reliability factor in the operation data of building-related energy systems. The data sanitation process includes the duty cycle (operational) detection, data validation with first principles, outliers detection, and data imputation. The machine learning algorithms and innovative visualizations are used to perform different tasks proposed in the work-flow. The focus of the data sanitation work-flow is to sanitize the energy-related monitoring data of buildings, with low configuration effort requirement, and domain knowledge for data analysis. The work-flow can be seen in the Figure 4.1. The work-flow is created to be generic and applicable to different types of energy systems data (time series data). The work-flow as can be seen in Figure 4.1, starts with data accessibility, followed by the automated detection of duty cycles (k-means), then detection of outliers in the data, subsequently followed by data imputation (sanitize gaps), and finally use visualizations (process limits) and first principles for validation.

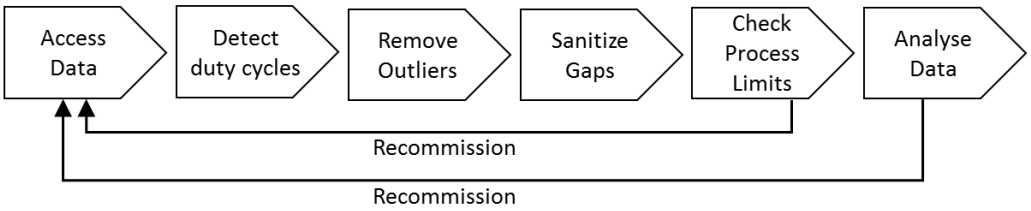


Figure 4.1: Data Sanitation Workflow

In the first section, different types of duty cycle detection algorithms are discussed, such as duty cycle detection with supervised learning (where already labeled data is required for training of model and in this research the electrical parameters were used for generating the labeled data for modeling), and duty cycle detection with unsupervised learning (k-means clustering). The subsequent section discusses the different first principles for data validation. The suggested first principle included to check whether the data are physically plausible, variables are correlated, rules using electrical parameters, using basic thermodynamics laws, and process limits check (testing if the values are in the process limit). Afterward, a method for finding the outliers in data, is

proposed in this research using the duty cycle (On/Off) based z-score with clustering. The duty cycle detection information is obtained using k-means for the proposed outliers detection method. At the end, data imputation with interpolation and regression is discussed. Furthermore, a data imputation method for energy systems in building is also proposed. The concept discussed in this chapter are published in [157, 158, 159, 160].

4.1 Duty Cycle (On/Off) detection

The automatic detection of duty cycles (operational) in the energy systems can be very useful in the analysis phase. In this section the different methods for automatic detection of duty cycle (operational) has been discussed depending upon the features available in different systems.

4.1.1 Methodology

In this section the methodology for the duty cycle detection has been discussed. There are different methods elaborated for automatic duty cycle detection either by selecting some features on the basis of knowledge or using supervised and unsupervised learning. For supervised learning various techniques have been tested i.g K nearest neighbor (KNN), support vector machines, and random forest, whereas k-means clustering algorithm has been selected in the category of unsupervised learning algorithms.

4.1.1.1 On/Off state detection using supervised learning

In order to detect the On/Off state of machine, it is suggested to consider the state of machine as On whenever there is electricity consumption by pumps in the given system. Figure 4.2 shows the 3D graph of the volume flows and energy in case of considering the electrical parameters for operational cycle detection. It can be observed from Figure 4.2 that there is either no or very less flows during the Off state in the high temperature (HT), medium temperature (MT) and low temperature (LT) cycle. Similar kind of behavior can be observed in the 3D graph of energies as well where it can be seen that there are no energy flows during Off state.

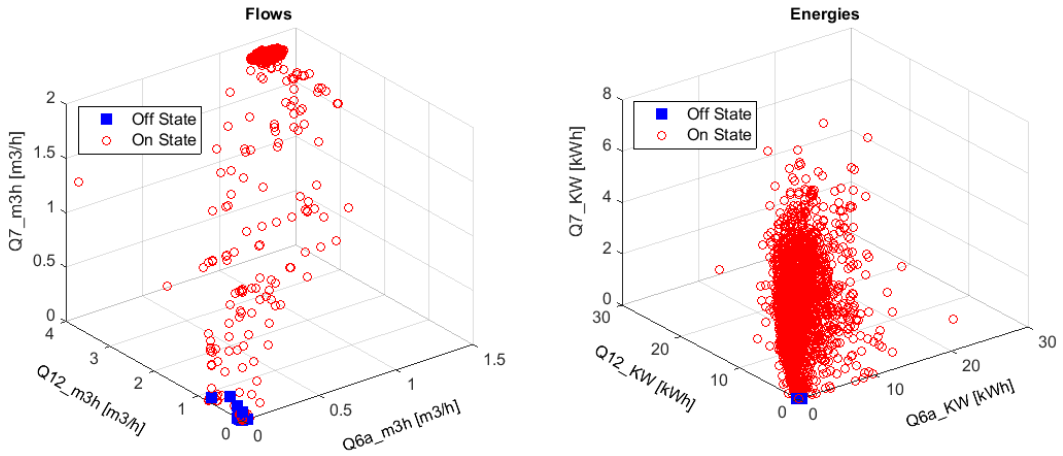


Figure 4.2: On/Off state based on electrical consumption of pumps in the chiller

In order to see in detail, the behavior of the three cycles can clearly indicate the On/Off working status of the energy system (chiller) as can be seen in Figure 4.3 with electricity consumption parameters of the pumps. The last graph in the Figure 4.3 shows the On/Off status where the zero value represents the off state and value greater than zero represents the On state of chiller. With each cycle the corresponding temperatures of low temperature cycle (LT), medium temperature (MT) and the high temperature cycle are displayed in the first three graphs. It can observed that during the On cycle the temperatures in High temperature (HT) cycle and Medium temperature (MT) cycle increases, whereas, at the same time decrease in temperature can be noticed in Low Temperature cycle. So the data can be labeled with On and Off state using

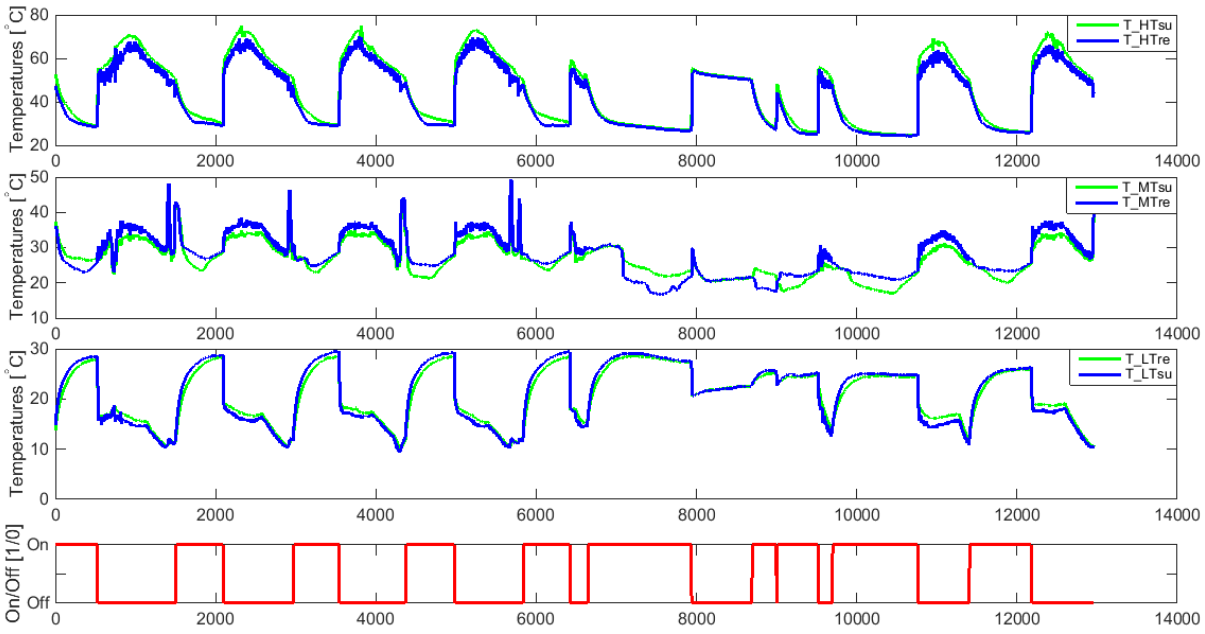


Figure 4.3: On/Off state based on electrical consumption of pumps in the chiller

the electrical parameters which will be used for supervised learning techniques. The consecutive On states will be considered as On (operational) cycle and similarly for the Off cycle.

For automatic detection of cycle with supervised learning, three methods are used as they belong to different approaches of supervised learning. The methods are as following

- **K Nearest Neighbors (KNN) Classification Algorithm**: In this algorithm the classification is done on the basis of the distance from the neighbors and the number of neighbors. A new object is classified by the majority vote of its neighbors that is the object is assigned to the class that is most common among its k nearest neighbors.
- **Support Vector Machine (SVM)**: A SVM algorithm is supervised algorithm that works for binary classes but there are methods available to make it work for multi class as well. The benefit of using SVM over perceptron is that it works even for data that is non separable. It is expensive in terms of memory and processing time.
- **Random Forest Tree Algorithm**: Random Forest is a tree based algorithm which makes many trees for classification. An object is assigned the class that is voted by the majority of the trees.

4.1.1.2 On/Off state detection using unsupervised learning

The previous method discussed uses the electricity consumption meter but there are other systems where electricity meters are not installed, therefore applying the supervised method is not possible in that case. So we need a robust method for duty cycle detection. If we analyze the behavior of the different temperatures, flows, energy readings and pressures, we can see that the data has two different patterns showing the On and Off state of machine as can be observed from Figure 4.4. Therefore, on the basis of this assumption that the machine will perform differently in two states we can cluster the data in two clusters using k-means clustering algorithm.

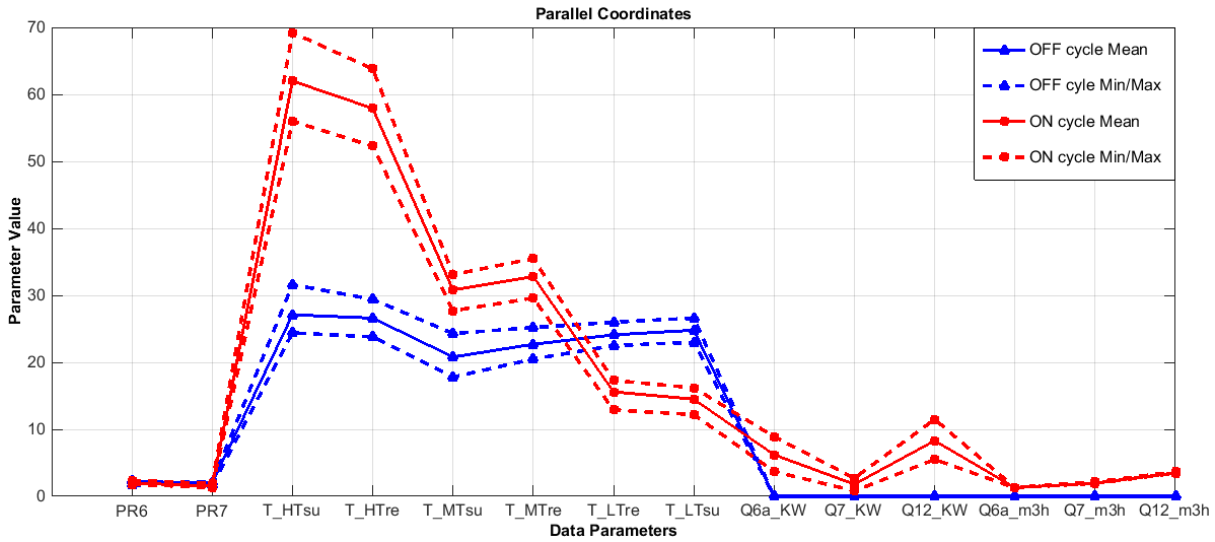


Figure 4.4: Parallel Coordinates showing Maximum, Minimum and Mean of On and Off Cycle

4.1.2 Experiments and Results

This section discusses the various experimental results performed for the duty cycle detection of the energy system. In the first results from the supervised learning that includes K nearest neighbor (KNN), support vector machines (SVM) and random forest are elaborated. Furthermore the results of k-means clustering are discussed that fall under the category of unsupervised learning.

4.1.2.1 Supervised learning

The experiments carried out for finding the duty cycle (On/Off) state classification with the three different supervised learning algorithm with different parameters of each algorithm and best results are reported. In each method a 10 fold cross validation [161] method is used where 90% of data is used for training and 10% data is used for testing. The three techniques are also tested with z-score normalization and with random missing values in order to check its sensitivity to missing gaps. The missing gaps are filled with linear interpolation and Nearest Neighbor (NN) method and are tested for accuracy of classification in each method.

a. On/Off State detection with KNN Algorithm

In order to achieve good results, experiments have been performed with different parameters of KNN algorithm like distance function (Euclidean and correlation) and value of K. The Figure 4.5 shows the classification accuracy with KNN algorithm. The line with diamond symbols is representing the classification accuracy of Off state data, whereas, the second line with rectangle symbol represents the On state classification accuracy rate. Along the x-axis the different methods are given. The first index in the x-axis shows the classification accuracy rate of the original data, the second shows the accuracy results with Z-Score, the third index in x-axis explain the results with missing values, whereas the last two show the results of classification accuracy of missing data filled with linear interpolation and nearest neighbor interpolation. The results have shown that KNN has given best result with K=5 and with distance function selected as Euclidean distance as can be noticed in Figure 4.5. The Z-Score has not shown much difference to the classification results. It is also clear from the Figure 4.5 that the missing values affect the classification results of KNN at around 79%.

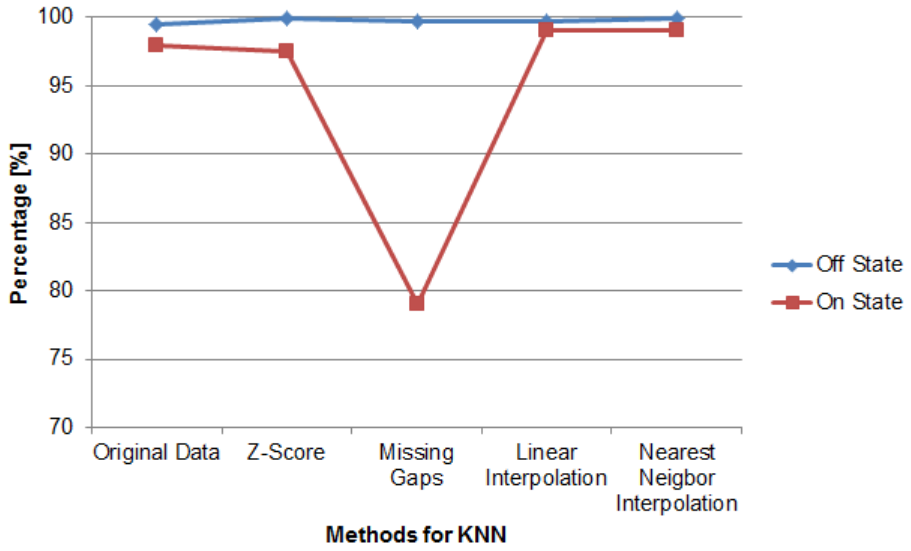


Figure 4.5: KNN Algorithm Classification with original data, Z-Score, Missing Value, Filled gaps with Linear and Nearest Neighbor Interpolation

b. On/Off State detection with Random Forest Tree Algorithm

The experiments have been performed using Random Forest Tree algorithm with different parameters. The Figure 4.6 shows the classification accuracy with Random Forest tree algorithm. The first column in each group shows the classification of Off state data along with misclassified Off state, whereas, the second column represents the On state classification rate. The first group explains the classification rate of original data, the second group shows the results with Z-Score, the third group explains the results with missing values, whereas, the last two groups shows the results of classification of missing data filled with linear interpolation and nearest neighbor interpolation. The results have shown that Random Forest has given has shown a minor difference with Z-score and is not sensitive to missing values as well. Furthermore, the data imputation with Nearest Neighbor (NN) has shown a slight better results than linear interpolation.

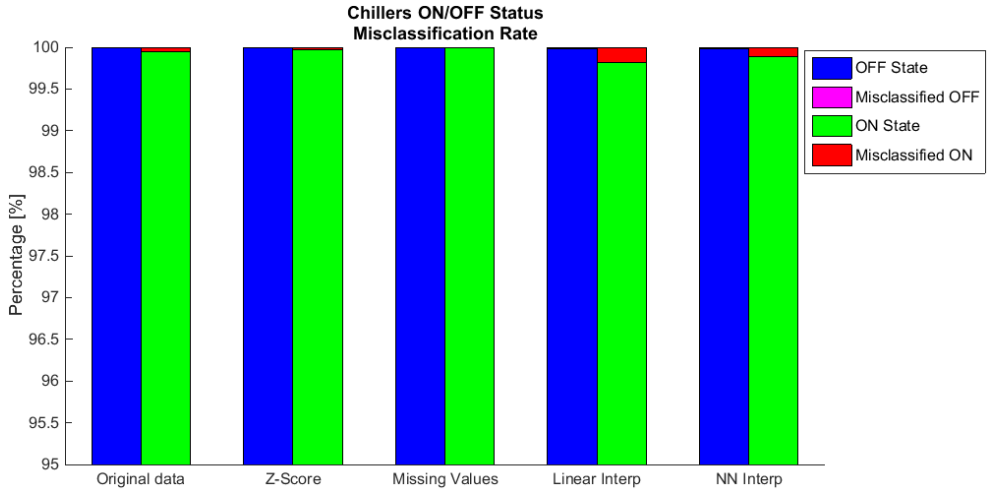


Figure 4.6: Random Forest Algorithm Classification with original data, Z-Score, Missing Value, Filled gaps with Linear and Nearest Neighbor Interpolation

c. On/Off State detection with Support Vector Machines (SVM) Algorithm

The different kernel functions (Linear and Radial Basis Function (RBF) were used to experiment with the support vector machines method. The Figure 4.7 shows the classification accuracy SVM algorithm with Linear and Radial Basis Function (RBF) kernels. The first column in each group shows the classification of Off state data along with misclassified Off state, whereas, the second column represents the On state classification rate. The first group (*Orig_Data_L*) explains the

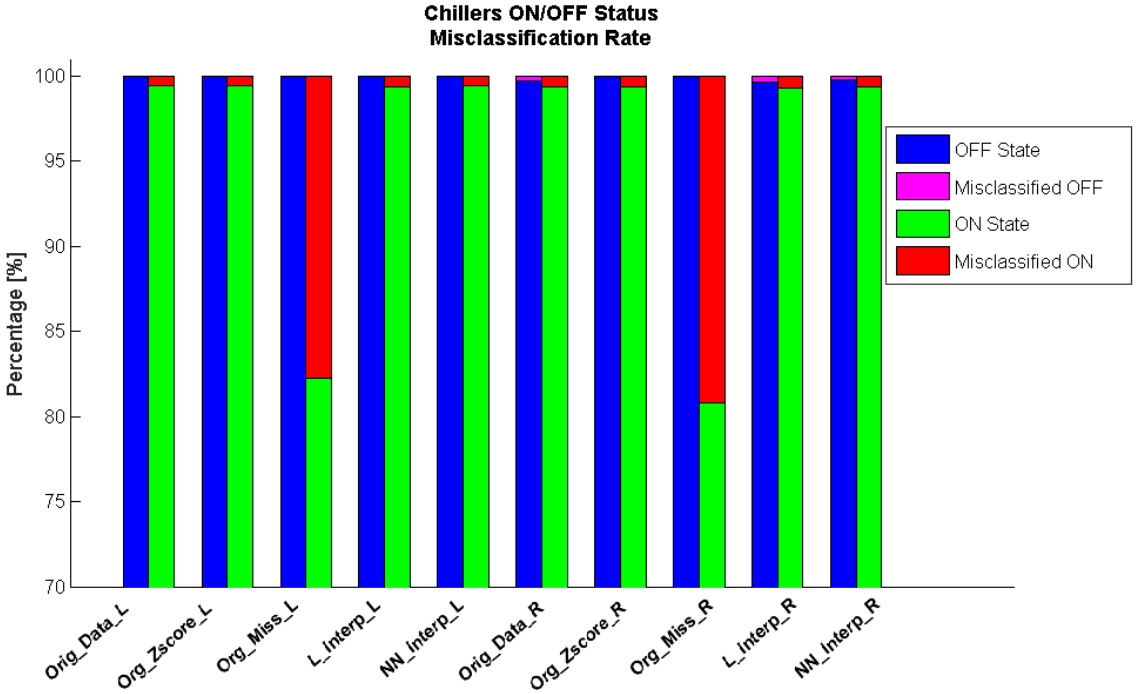


Figure 4.7: SVM Algorithm Classification with original data (Linear and RBF Kernels), Z-Score, Missing Value, Filled gaps with Linear and Nearest Neighbor Interpolation

classification rate of original data with linear kernel, the second group (Org_Zscore_L) shows the results with Z-Score with linear kernel, the third group (Org_Miss_L) explains the results with missing values using linear kernel whereas the next two groups (L_Interp_L and NN_Interp_L) shows the results of classification of missing data filled with linear interpolation and nearest neighbor interpolation using linear kernel. The same process is repeated using Radial Basis Kernel (RBF) which are represented with $Orig\_Data\_R$, Org_Zscore_R,Org_Miss_R, L_Interp_R and NN_Interp_R as can be seen in Figure 4.7. The results have shown that Random Forest tree has given best result with RBF kernel as can be noticed in Figure 4.7. The Z-Score has not shown much difference to the classification results. It is also clear from both the Figure 4.7 that the missing values affect the classification results.

The Figure 4.8 shows the comparison of the three supervised learning used. The random forest has shown the best results among the three methods as it is also not sensitive to the missing values and overall results are quite satisfactory as compared to the other two methods. The KNN has shown that it is sensitive to noise with the lowest accuracy of around 90% among all the methods, but has shown good results when values were filled with interpolation method. The support vector machine has also shown sensitivity to noise but has performed better than KNN.

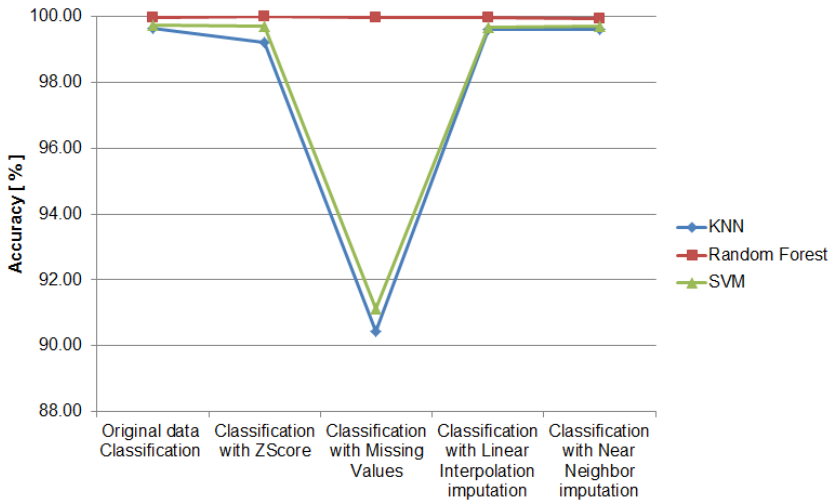


Figure 4.8: Comparison of the three supervised classification techniques

4.1.2.2 Unsupervised Learning

As discussed in the methodology section of the chapter the data varieties in two state On and Off, therefore k-means clustering algorithm has been used to find the two clusters (On/Off) in the data. Before applying the k-means algorithm the data is normalized using Min-Max normalization. The number of cluster for k-means clustering is taken as 2 with the Euclidean distance function, as the reason is the energy system behavior will vary between two states i.e. On and Off state, thus k-means will be able to detect these two states. The Figure 4.9 given below, demonstrates the On/Off status through K-means clustering algorithm that is applied. The x-axis is representing the time in minutes for all graphs in Figure 4.9.The first graph in Figure 4.9 shows the high temperature (HT) cycle temperatures, while, the second graph display the medium temperature (MT) cycle temperatures, whereas, the third graph represents the temperature of low temperature (LT) cycle, and the last graph demonstrates the On and Off status of the chiller.

It can be observed from the Figure 4.9 that the behavior of the temperatures of the LT, MT and HT, which is quite similar to the supervised learning method as discussed. Furthermore, the behaviour of the temperatures at the Low Temperature (LT), Medium Temperature [MT] and High Temperature [HT] cycle are responding according to the detected On/Off state. The dotted rectangle shows one On cycle of the chiller. It is evident from the Figure 4.9 that as per the operational pattern of chiller, during the detected On cycle, the LT temperatures decreases presenting the cooling operation. While, at the same time, the temperatures increase in the HT and MT cycle of the chiller. This change pattern in temperature, is a clear signal that the chiller is in operational mode, also detected by k-means clustering which can be seen in Figure 4.9.

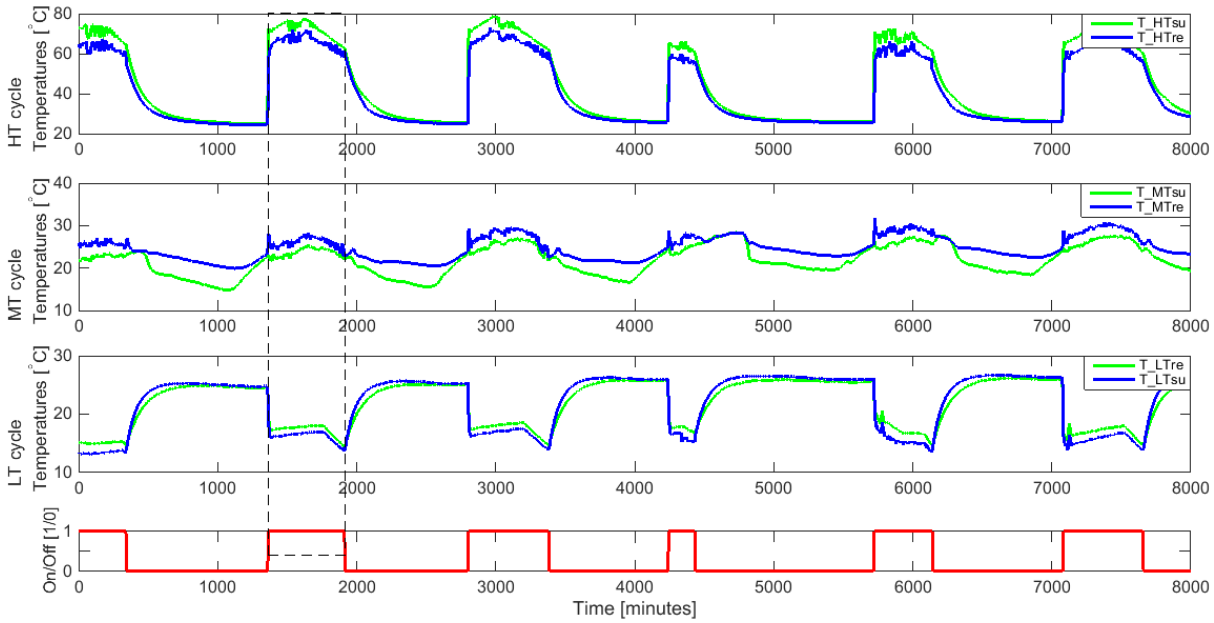


Figure 4.9: On/Off Cycle Detection Using K-Means Clustering Algorithm in Chiller data

Furthermore, the k-means clustering algorithm is also used for finding the presence of the occupants in the building, using the air quality information (temperature, relative humidity and carbon dioxide) of a ventilation system in building. The ventilation system that is analyzed uses the different sensors for calculating the temperature, humidity and carbon dioxide in the building. These values are used to handle the ventilation systems for maintaining good air quality in the building. The total number of used sensors data is 36 for automatic detection of the presence of occupants in the building. The idea is that with the occupants in the building the behavior recorded in these sensors will differentiate from the patterns of the building when there are no occupants. The data is normalized using min-max normalization and then k-means is used with euclidean distance and two number of clusters (yes for occupants in the building and no for without occupants). The automatic detection of occupants can help in efficiently managing the ventilation system and lightening system in the building as well, which will help in saving energy. The Figure 4.10 shows the behavior of the sensors of air quality inside the building. The x-axis in Figure 4.10 is representing the time in minutes, and it can seen that the indoor environment in the building is maintained as the temperature range is between $21°C$ and $24°C$, the relative humidity is kept around 38% during the presence of occupants in the building, and carbon dioxide level below $1500ppm$. The first graph in the Figure 4.10 shows the temperature sensors placed at different locations in the building, while the second graph is representing the relative humidity

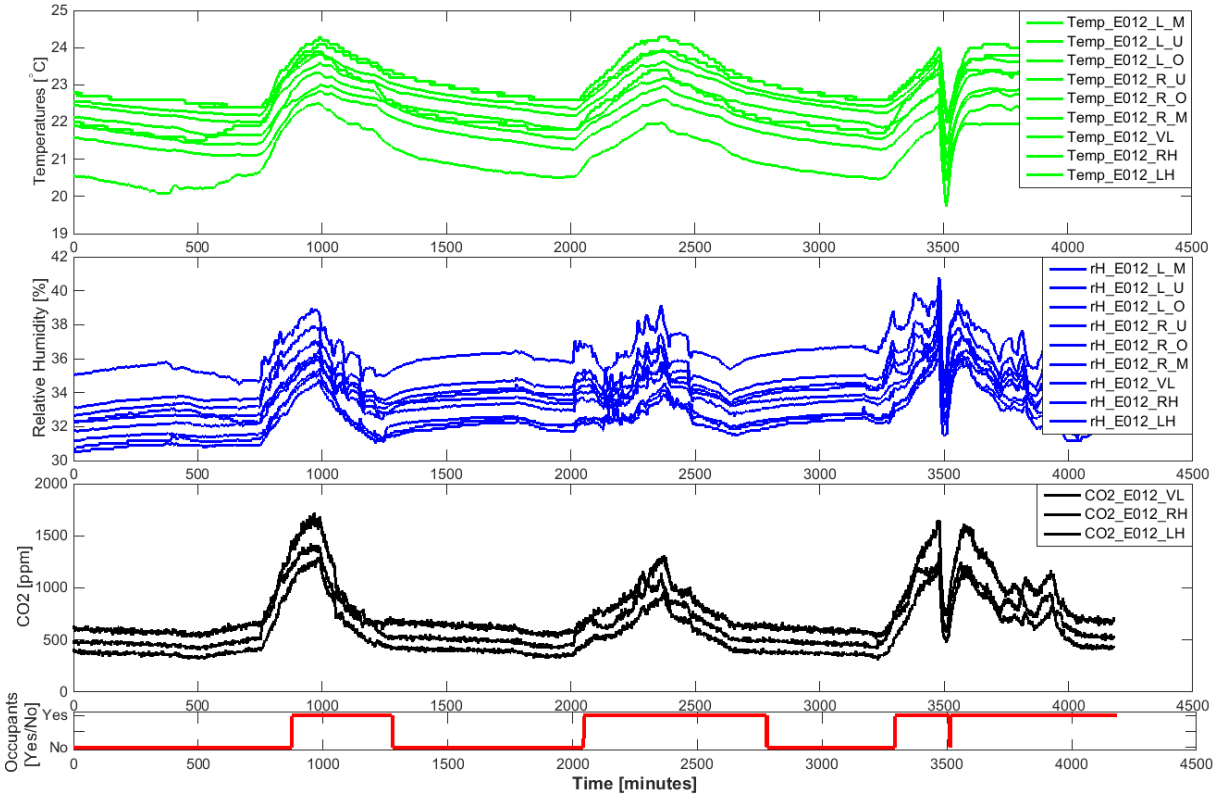


Figure 4.10: Occupant's Presence Detection Using K-Means Clustering Algorithm in Ventilation data

in the building, and the third graph displays the carbon dioxide level maintained in the building. The last graph in the Figure 4.10 shows the k-means results for finding automatically the occupants in the building. When the quality of air decreases, it shows that there are occupants in the building and ventilation system running at these points for maintaining good air quality in the building. Similarly the quality of air remains good while there are no occupants in the building and it has been observed that during air quality remains good during the night time and on weekends, thus ventilation system does not need to run in this period. It is noteworthy to observe that there is a sudden change in values at around 3500 minutes in the Figure 4.10. It has been observed that this type of behavior occurs when the window is opened by the occupants in the building and that's why from the air quality perspective the k-means have detected it as no occupant in building for around 15 minutes.

4.2 Data Validation with First Principles

In this section various first principles and visualization are suggested for validation of data. The validation of data is required, so that it should be confirmed that the data exhibits the required behavior before detail analysis of the data. The data validation step will add reliability to the data, thus will help to make a first impression of the data quality before detailed analysis of the energy systems in buildings. There are different methods that can be used as a validation such as first principles, process limit checks, performance indicators, and different visualizations.

Data Validation using Energy Balance

In case if the system boundaries are known and all relevant parameters are available then the first law of thermodynamics can be used as first principle for detection of faults in the data using the energy readings. The system on which the experimentation has done has three heat meters measuring the energy flow in the low, medium and high temperature cycle, thus allowing to define a system boundary. The only parameter unknown is the energy losses in the system. According to the law of thermodynamics the energy that flows into the system should be equal to the energy that flows out of the system given by Equation 4.1

$$Q_{LT} + Q_{HT} - Q_{MT} + \Delta E = 0 \tag{4.1}$$

Therefore,

$$\Delta E = Q_{MT} - Q_{LT} - Q_{HT} \tag{4.2}$$

Here ΔE represents the energy losses in the system. During the normal operation of chillers, it is assumed that ΔE should be near to 0. The Figure 4.11 shows the ΔE for eight On duty cycles

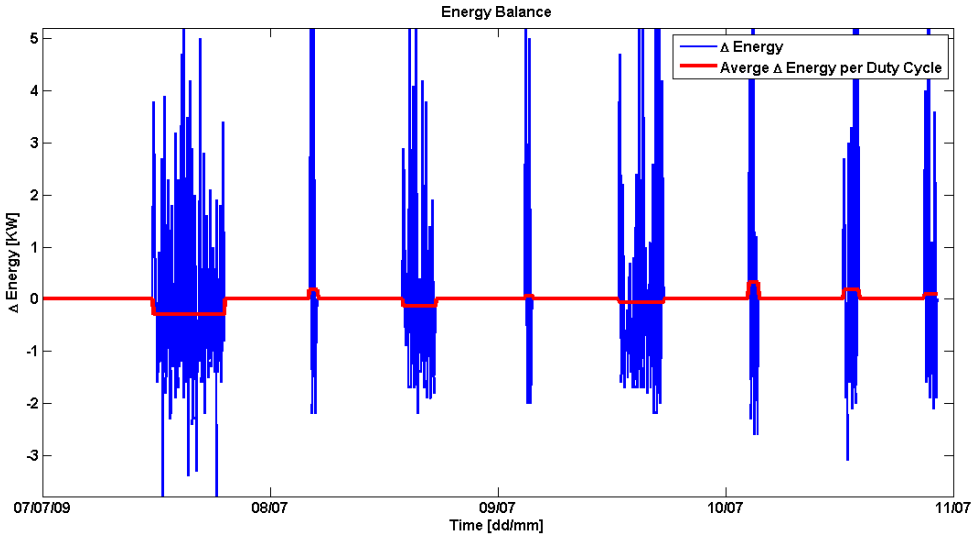


Figure 4.11: Average Energy Balance per duty cycle

of the adsorption chiller. Due to the various delays in thermal dissipation and mass flows the energy balance on a sample-by-sample base shows very high values which seems unacceptable, which can be seen in Figure 4.11 (blue line). Thus it is suggested to use the average ΔE of each duty cycle for data validation. In Figure 4.11, the (red graph) ΔE ranges from -0.2 to 0.2 kW, which is an acceptable range. Therefore the energy balance is plausible and the data can be used for further analysis. The automated detection of the duty cycle helps to do this analysis as the data in Off-state are commonly not relevant.

Data Validation Using Coefficient of Performance (CoP)

The coefficient of performance (CoP) can be used as a good indicator for validating the energy system data, as with every system, the CoP for normal operation is given, and the out of range

CoP values can be used as an indicator for some data anomalies. The thermal CoP range of the solar based cooling system is between 0 to 1. Similarly the Electrical CoP range depends on system. Both CoP values should be between the threshold values for validated data. The equations for find CoP thermal and CoP electrical are as following Equations 4.3 and 4.4

$$CoP_{TH} = \frac{Q7}{Q6a} \tag{4.3}$$

$$CoP_{ELEC} = \frac{Q6a}{Total\ Energy\ Consumed} \tag{4.4}$$

Due to various problem such sensor calibration issues, there is a high chance that the value of the CoP does not fall in range. In order to address this issue, it is suggested in this research to find average CoP of one duty cycle and data validation is done on the basis of it.The Figure 4.12 shows the thermal CoP on a sample-by-sample base (blue graph) and averaged over the On-state of a duty cycle (red graph). The average value of CoP is around 0.5 during the On cycle that gives a good signal that the data is plausible.

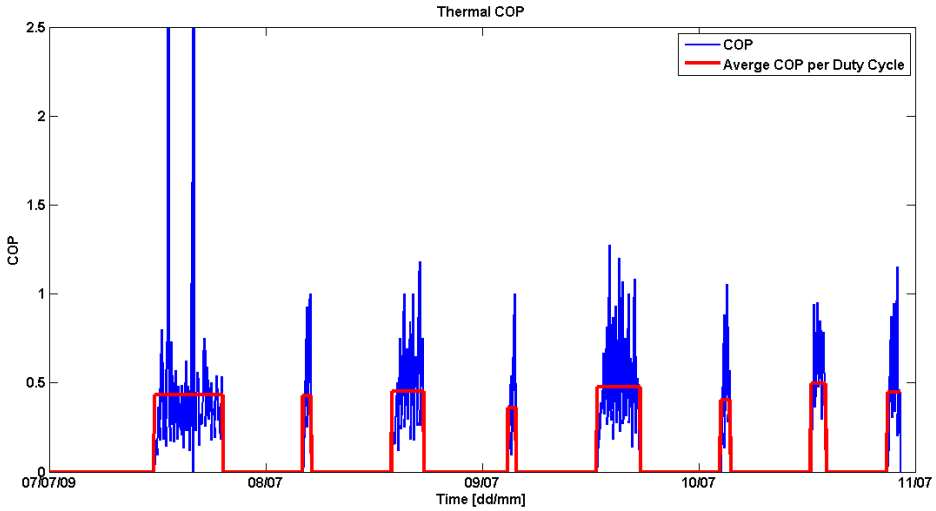


Figure 4.12: Sample by Sample CoP and average CoP of each cycle

Data Validation using temperature sensor values during On cycle

According to the first law of thermodynamics, the heat flows from upper temperature towards lower temperature. Thus the temperature of heating water flow coming inside the chiller from HT cycle should be greater than the temperature going out during the On (operational) cycle. Therefore relationship for temperature parameters in high temperature circuit in On cycle can be defined as follows in Equation 4.5.

$$HT_{su} > HT_{re} \tag{4.5}$$

Similarly for cooling side the temperature of chilling water flow leaving chiller should be lesser than return flow. Therefore temperature parameters in low temperature circuit during the On cycle can be defined as in Equation 4.6.

$$LT_{re} > LT_{su} \tag{4.6}$$

Also the temperature of cooling water flow to the cooling tower should be higher than the return flow temperature as heat is released in the cooling tower. Therefore related temperature condition for the medium temperature circuit during the On (operational) cycle can be defined as Equation 4.7

$$MT_{su} > MT_{re} \tag{4.7}$$

Thus violation of above external flow temperature conditions can be acknowledged as an implausible data.

Data Validation Using Electricity Consumption of pumps parameters

First principles can be very handy for a quick data validation. In this section some first principles are proposed, while using the knowledge of the system. The first step done in this subsection is to explain some functions which are later used in the following subsection to define the various types of anomalies in the data. The Is_On function checks if the given module(Pump) is on. The input to the function are the volume flow and electricity consumption of the pump. In the case of a pump it will check if the pump is consuming electricity and also showing volume flow. It will return 1 if electricity consumption is greater than 0 or there is volume flow. It can be declared as using the Algorithm 4.1.

Algorithm 4.1: Is_On Pseudo Code

```
Input: Pump_flow, Pump_electricity_consumed;
Initialize: status := 0
begin
if (Pump_flow > 0) AND (Pump_electricity_consumed > 0) then
    status := 1
end
return status
end
```

The *Find_Performance* function will find Performance of Pump (PoP) with flow and electricity consumption parameter, as explained in Algorithm 4.2. The function $Find\_Performance$ will return the calculated performance of the pump.

Algorithm 4.2: Find_Performance Pseudo Code

```
Input: flow, electricity_consumed;
Initialize: PoP := 0
begin
PoP := flow / electricity_consumed
return PoP
end
```

The *Find_cycle* function as defined in Algorithm 4.3, will find the length of the operational duty cycle's in the given data. The input to the function are the data and time stamp. It will return *cycle_time* that will be the time when the value consists of consecutive non-zero values of either electricity consumption or flow parameter, whichever is passed to function.

Algorithm 4.3: Find_Cycle Pseudo Code

```
Input: Time_Stamp, data;
Initialize: cycle_time := 0, start := 0, end := 0, flag := 0
begin
for i = 1 : size_of_data do
    if data[i] > 0 AND flag = 0 then
        start := i
        flag := 1
        Else
        end := i
        flag := 0
    end
    cycle_time := Time_Stamp[end] − Time_Stamp[start]
end
return cycle_time
end
```

The following are the different first principle tests proposed for validation of data using the electrical parameters of the system,

1. The maximum and minimum length of the cycle is checked whether it is too short or too long. The *Find_cycle* function is used to calculate the time length for the given data.The too short or too long cycles like the On (operational) cycle of more than 18 hours will be considered as implausible such as in case of solar based energy systems. The Algorithm 4.4 defines the rule for the discussed rule. The inputs to the Algorithm 4.4 are *data*, *threshold_time* and it returns the *check* value as binary.

Algorithm 4.4: Find_operational_cycle_length Pseudo Code

```
Input: data, threshold_time;
Initialize: check := 0
begin
if (Find_cycle(data) < threshold_time) OR (Find_cycle(data) > threshold_time) then
    check := 1
end
return check
end
```

2. If the chiller is On, all the pumps should be working which means that they should be consuming electricity and showing volume flow which can be seen in Algorithm 4.5. The input to the Algorithm 4.5 are electricity consumption parameters ($E6$, $E7$, $E8$) and volume flow ($Q6a\_m3h$, $Q12\_m3h$, $Q7\_m3h$) and it will return the data $Mark\_Error$ where the rule has been violated.

3. The point, where the performance of the pump (PoP) is either too low or high requires to be validated, as normally the pumps work in the specific range. It can also be helpful in

Algorithm 4.5: Pump test with Electricity_Volume Pseudo Code

```
Input: E6, Q6a_m3h, Q7_m3h, E8, Q12_m3h, E7, Status_chiller;
Initialize: Mark_Error := 0
begin
for i = 1 : size_of_data do
    if Is_On(Q7_m3h[i], E8[i]) = 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
    if Is_On(Q6a_m3h[i], E6[i]) = 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
    if Is_On(Q12_m3h[i], E7[i]) = 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
    if Is_On(Q7_m3h[i], E8[i]) = 1 AND Status_chiller[i] = 0 then
        Mark_Error[i] := 1
    end
    if Is_On(Q6a_m3h[i], E6[i]) = 1 AND Status_chiller[i] = 0 then
        Mark_Error[i] := 1
    end
    if Is_On(Q12_m3h[i], E7[i]) = 1 AND Status_chiller[i] = 0 then
        Mark_Error[i] := 1
    end
end
return Mark_Error
end
```

finding faults in the pumps. The following check can be seen in Algorithm 4.6 will find the places where the performance of the pump in the three cycles (HT, LT, MT) is below the required value. The Find_Performance function is used to find it and the input to the function will be the energy consumption and volume flow in each cycle.

Algorithm 4.6: Find_operational_cycle_length Pseudo Code

```
Input: flow, electricity(Elec), lower threshold (σ1), upper threshold (σ2);
Initialize: check := 0
begin
if (Find_Performance(flow, Elec) < σ1) OR (Find_Performance(flow, Elec) > σ2) then
    check := 1
end
return check
end
```

4. To check that when a pump is consuming electricity then it should also show volume flow. Otherwise if the pump is using electricity and there is no volume flow then it is not possible in a normal case. The same applies for the volume flow, if there is volume flow and no

energy is consumed then not a normal case as can be observed in Algorithm 4.7. The input to the function are the electricity consumption and volume flows in the high temperature (HT), medium temperature (MT), and low temperature (LT) cycle with the duty cycle information. The output of the function will be $Mark$ that will show the instances where the rule has been violated.

Algorithm 4.7: Electricity_Volume test Pseudo Code

```
Input: E6, Q6a_m3h, Q7_m3h, E8, Q12_m3h, E7, Status_chiller;
Initialize: Mark_Error := 0
begin
for i = 1 : size_of_data do
    if Q7_m3h[i] > 0 AND E8[i] = 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
    if Q6a_m3h[i] > 0 AND E6[i] = 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
    if Q12_m3h[i] > 0 AND E7[i] = 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
    if Q7_m3h[i] = 0 AND E8[i]) > 0 AND Status_chiller[i] = 0 then
        Mark_Error[i] := 1
    end
    if Q6a_m3h[i] = 0 AND E6[i] > 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
    if Q12_m3h[i] = 0 AND E7[i] > 0 AND Status_chiller[i] = 1 then
        Mark_Error[i] := 1
    end
end
return Mark_Error
end
```

5. In the last rule equation for finding the energy from temperature difference and flow. The energies of each cycle can be calculated by using the temperatures and flows variables. The equation shows the relationship between these variables. We can validate the value of each variable against the other given variables. In the equations discussed LT cycle have taken considered. The same process can be applied for the other two cycles as well. The calculated values can be cross check with the recorded energy values of the respective cycle, thus will help in validation of the energy values. The Equation 4.8 can be used for conversion which is given as following,

$$Q = \Delta T.C_P.m \tag{4.8}$$

where $\Delta T = \mid T\_LTre - T\_LTsu \mid$,
m is the mass,
and C_P is the specific heat capacity of refrigerant.

Data Validation with Correlation of the parameters

One of the methods for data validation can be the correlation between the different parameters taken under consideration for the research during the operational cycles of the energy system. The Figure 4.13 shows the correlation between parameters where the number inside each block shows the correlation value. The positive number (0 to 1) shows the positive correlation with 0 as no correlation and 1 as maximum correlation. Similarly the negative number shows the negative correlation. The On cycles data have been taken for the Figure 4.13 . It can be observed from the Figure 4.13 that all parameters are positively correlated except the temperature variables in Low temperature cycle as during the On cycle the temperatures in low cycles decrease while the other parameters increases, thus shows that the data is plausible. The graphs at the diagonal positions in Figure 4.13 are representing the histogram of each parameter, whereas, the non-diagonal positions show the scatter plot, and the line inside the non-diagonal graphs (scatter plot) denotes the slopes of the least-squares.

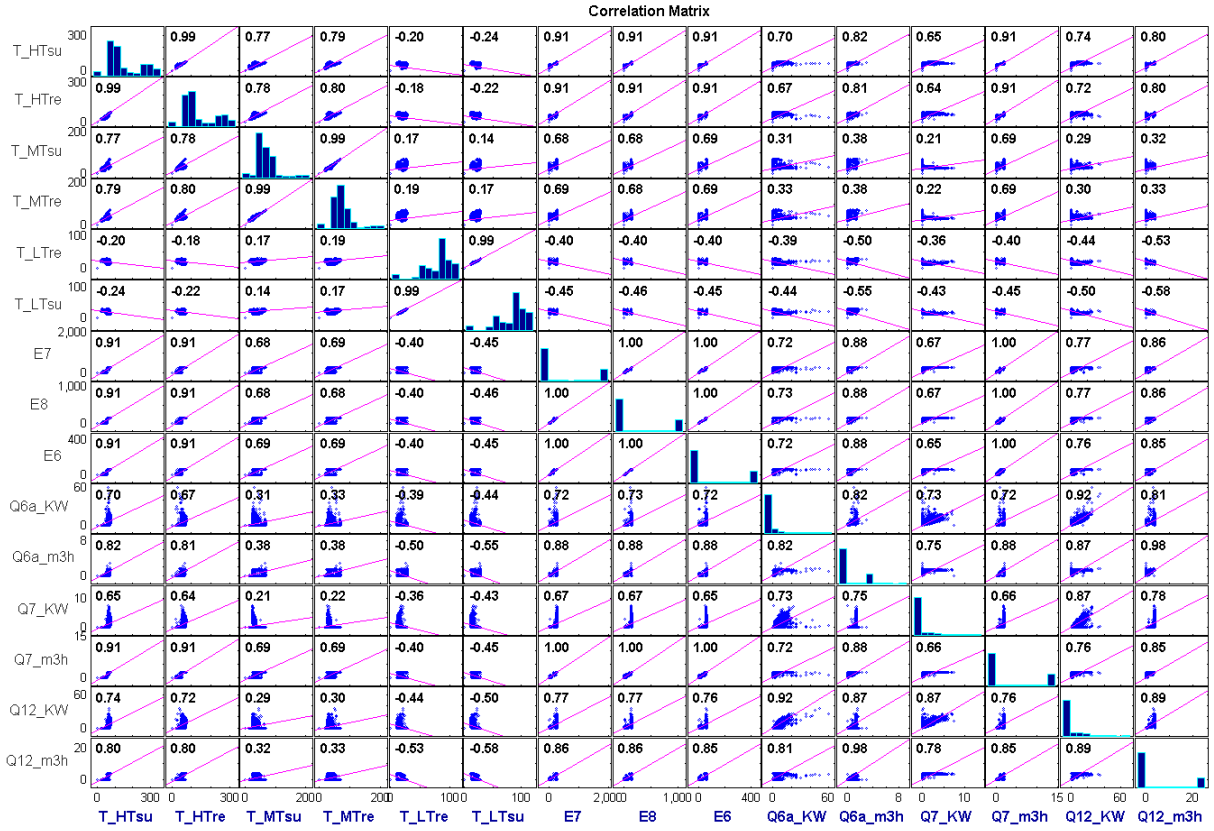
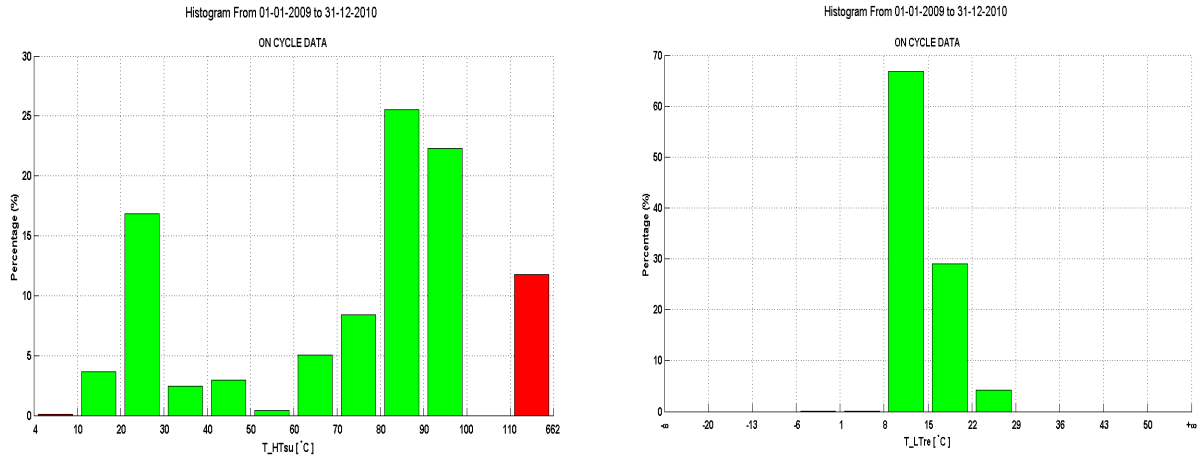


Figure 4.13: Correlation of parameters

Data Validation with Range threshold of each parameter

The check against the range limits can be used to visualize statistical information about the data distribution during the On cycle of the system. This will help in providing a instantaneous way of visually checking data plausibility. The visualization of only On cycle data will significantly reduce the number of process limit violations, thus giving more understanding into the real operation

of the system. For Figure 4.14(a) and Figure 4.14(b) the data within the range limit are equally divided into ten bins and two additional range limits violation bin at upper and lower side that contains all the remaining outliers. Figure 4.14(a) shows the temperature at low temperature return side (LTre) contains data with no outliers, which is indicated by the violation bins to have $+\infty$ as the upper boundary and -∞ as the lower boundary. Figure 4.14(b) shows the temperature at High temperature supply side (LTre) contains data with outliers, which is indicated by the violation bins to have the maximum value ($662°C$) it has at the upper boundary and 0 as the lower boundary. The violation value bins are drawn with red bars for better understanding as can be noticed in Figure 4.14(b).



(a) Histogram of the return temperature in the low temperature cycle without the outliers

(b) Histogram of the supply temperature in the high temperature cycle along with the outliers

Figure 4.14: Histogram representation of data

4.3 Outliers Detection Using Cycle Based Z-Score Normalization and Expectation Maximization (EM) Clustering Algorithm

In this section a method has been proposed for outlier's detection in data of energy systems of the buildings. Outliers are the data points that does not exhibit the behavior of the desired process or system. These outliers are found in data that are usually added due to some noise in communication link, sensors or other reasons, while monitoring and saving the data. It is important to detect these outliers and replace them as it can effect in the detail analysis of the systems for faults detection and diagnosis. For handling the outliers in the data of energy systems in buildings a method has been proposed that uses the Expectation Maximization (EM) clustering algorithm, with the duty cycle (information extracted using the k-means clustering algorithm) based z-score normalized data.

4.3.1 Methodology

In this section the methodology for the proposed method for outliers detection has been discussed in detail. The method uses the duty cycle information extract while applying the k-means algorithm. The proposed technique uses the cycle based z-score for detection of outliers. The

problem with taking the normal z-score is that it considers the mean and standard deviation of the whole data for normalization, which might not make the outliers away from the normal behavior, thus makes it difficult for clustering algorithm to identify the outliers. It is very difficult to find these kinds of faults with normal normalization process as these points need to be more highlighted than the normal behavior of the data in that specific machine cycle. The benefit of highlighting these points will help the clustering algorithm to cluster these points away from the normal behavior of the data. Thus help in finding the invalid data points in the data.

Furthermore, as we are dealing with the solar cooling system, where the behavior of the system will vary during the On and off cycle. At the same time it may differ on a rainy day than the sunny day, therefore the mean of the sunny day will be higher than the rainy day as well the standard deviation. If we are considering the z-score for the whole data, there is a higher chance that we neglect or consider the correct data points as erroneous data points as the dominated data e.g. sunny days (high temperature) behavior will overcome the behavior of rainy days (low temperature).

To overcome this problem, it has been proposed in this research to use the Z-score as Equation 4.9 discussed as following

$$\mathrm{Z-Score_{Cycle}} = \frac{\mathrm{X} - \mu_{\mathrm{Cycle}}}{\sigma_{\mathrm{Cycle}}} \tag{4.9}$$

Where $Z - Score_{Cycle}$ is the z-score for each Cycle when the system is either On or Off, X is the value of the sensor,μ_{Cycle} is the population mean of each Cycle, and σ_{Cycle} is the standard deviation of each Cycle. Moreover, it can be noticed in Figure 4.15 that the data points at around

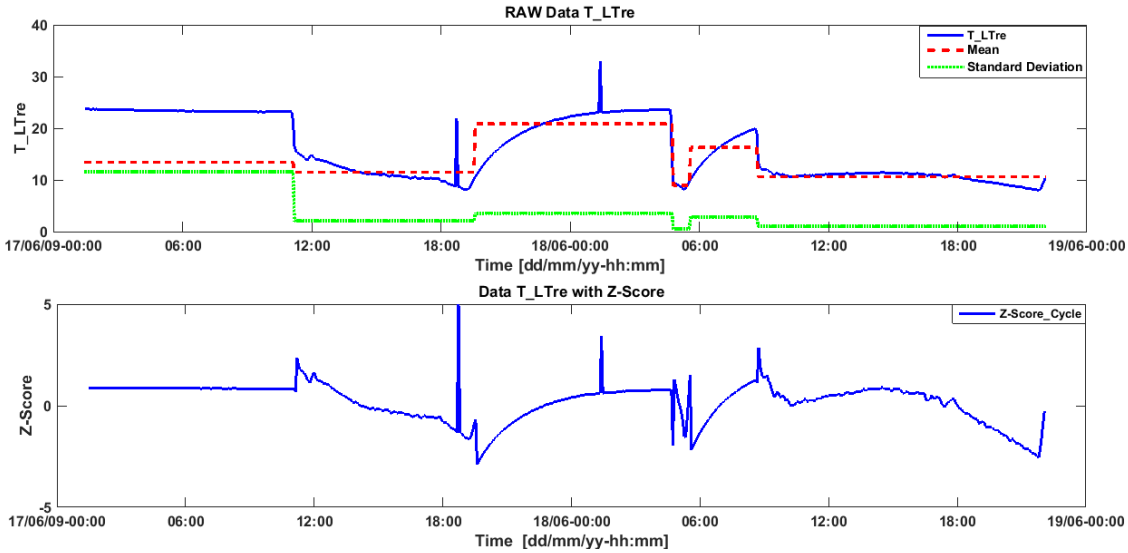


Figure 4.15: Raw data of low temperature at return side and cycle based Z-Score of the raw data

2009-06-17-18:00 and around 2009-06-18-00:00 are moved away from the mean of the normal data. The benefit of this will be while using the EM clustering algorithm as these points have been moved away from the normal behavior and will help it to cluster away from the normal behavior. There is also a problem in each cycle especially in the transient phase where the data is far away from the mean value of the cycle. To tackle this issue the initial 4 minutes in On state and 30 minutes in the off state are not considered for the mean and standard deviation of the cycle.

4.3.2 Experiments and Results

This section discusses the experiments and their results that are performed for the testing the proposed duty cycle based z-score outlier detection method. In order to detect outliers in the data the duty cycle based z-score is applied as described in Methodology Section of the topic. The statistical data for the z-scores (μ and σ) are derived based on the On and Off cycle instead of the whole data. Figure 4.16 shows the original data (T_LTre) along with the cycle based z-score; the two detected outliers (discussed in the methodology of the topic) as using expectation maximization clustering (EM) algorithm are marked with red circles both in the corresponding

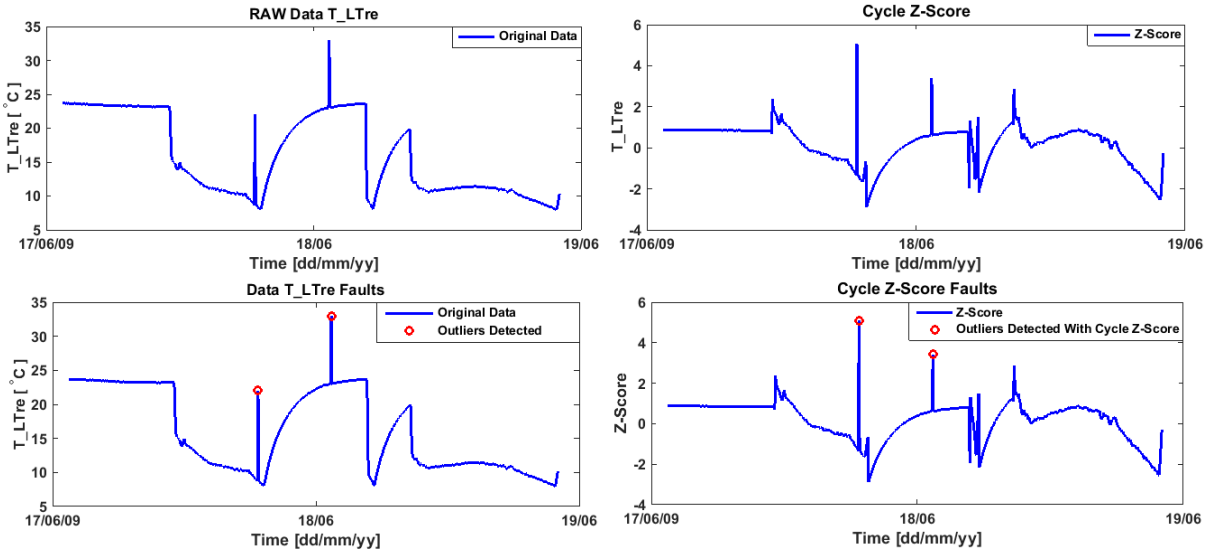


Figure 4.16: Outliers Detected with Cycle based Z-Score with EM Clustering

original data (T_LTre) and cycle based z-scored data. These outliers are, however, representing only single samples that can be reinstated by using linear interpolation for filling the missing gap as discussed in the next section.

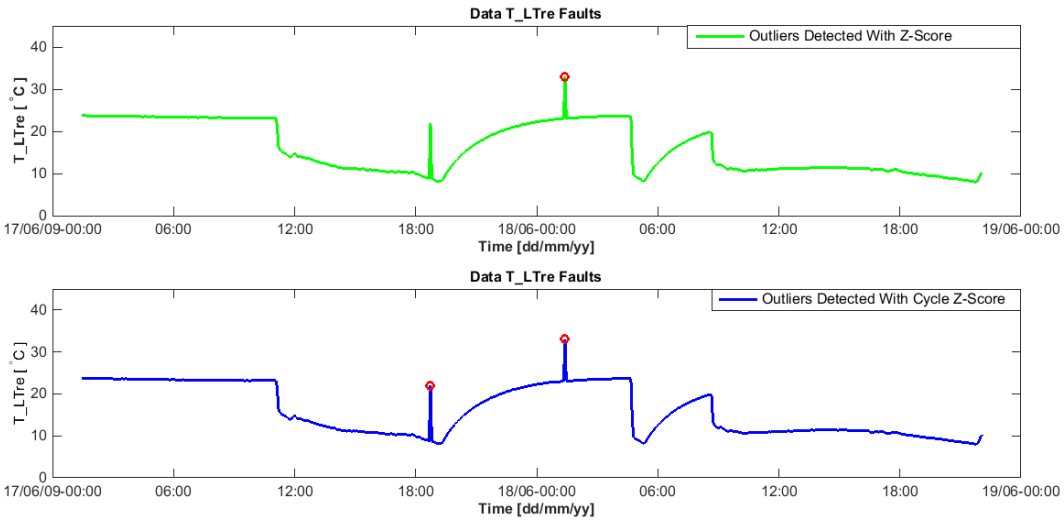


Figure 4.17: Comparison of Outliers detected with z-score and cycle based z-score

Furthermore the change of specific On or Off state based Z-Score parameters allows better performance for detection of outliers, as the behavior of a process varies strongly in the two different states. In Figure 4.17 the outliers have been detected with clustering using normal z-score and cycle based Z-Score. The first graph in Figure 4.17 represents the normal z-score based outlier detection where mean and standard deviation for the whole available data is used, whereas, the second graph shows the cycle based z-score where z-score is based on each On or Off cycle of the chiller. The graph has data with outliers detected with both methods represented with red circles. It can be seen clearly from the Figure 4.17 that the normal z-score has detected outliers that were away from mean of the whole data present at around 2009-06-18-0:00 and did not detected outlier at around 2009 06-17 at 18:00. But, cycle based z-score has detected the other outliers along with the neglected outlier, as the cycle based z-score makes the outlier detected at around 2009-06-17-18:00 away from mean according to the mean and standard deviation of that specific cycle. This helps the clustering algorithm to consider it as outlier.

4.4 Data Imputation with Interpolation and Regression

This section discussed the two methods i.e linear interpolation and regression, which have been used for data imputation of the missing data. There are missing gaps in the data due to various reasons such as communication failure or sensor failure etc. The decision of replacing these missing gaps (data imputation) is very critical as this can affect the analysis of any system. The simple most approach can be to ignore these parts where the data is not available. But the decision of ignoring the data is possible only when smaller portion of data is missing. There are other methods that can used for data imputation such as using mean, median, nearest neighbor, interpolation and regression depending upon the nature of the data. As there will be some missing gaps introduced because of the data validation and outlier detection step, therefore it is required to fill these gaps.

For testing the two methods, the monitoring samples are deleted randomly for filling gaps. The Figure 4.18 consists of four graphs i.e. the top left graph is representing the original data, while

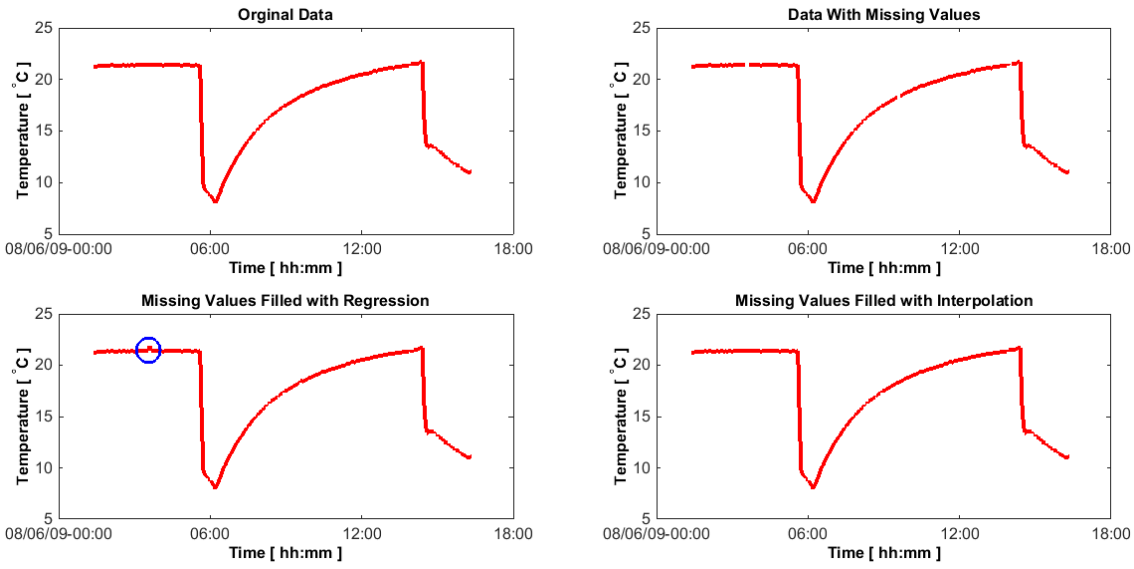


Figure 4.18: Filling shorter gaps of missing data with linear interpolation and regression

the top right graph shows the data after having short missing gaps, the bottom left is displaying the data where missing gaps are filled with regression, and the bottom right graph denotes the data where missing gaps were filled with the linear interpolation method. It is noteworthy to observe that the regression method adds some small peaks to the missing gaps as denoted by the blue circles in the Figure 4.18. On the other hand, interpolation has produced a smooth data in the case where there are non consecutive random missing values for short period of time as can be noticed in Figure 4.18. For comparison of the two methods used for missing gap imputation i.e regression and interpolation, an error graph has been drawn as can be seen in the Figure 4.19. The first graph in the Figure 4.19 shows the error (difference between the actual and predicted

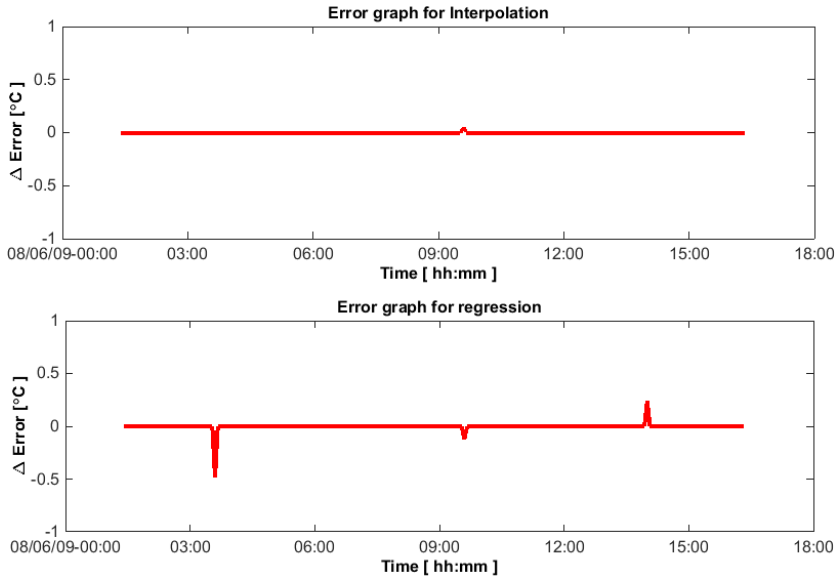


Figure 4.19: Error graph for shorter gaps with linear interpolation and regression

value) for the interpolation method, whereas, the second graph shows the error for the regression method. It can be observed from the Figure 4.19 that the error of interpolation is less as most of the error values are near 0

For testing the suggested data imputation methods against the long missing gaps, the data with large gap has been deleted randomly. The two methods i.e. interpolation and regression are tested for data imputation of long gaps. The Figure 4.20 shows the results for long gaps data imputation in data. The top left graph represent the original data, whereas, the top right display the data having the random long gap deleted part, the bottom left part represents the data having imputation with interpolation method, and the bottom right shows the data with imputation using regression method. Figure 4.20 shows that the regression method has shown better results as compared to interpolation.

Similarly an error graph has been drawn for comparing interpolation and regression imputation method for long missing gaps. The long gap has been introduced randomly for testing the two imputation methods. The first graph in the Figure 4.21 is representing the error graph for interpolation method ,while the second graph displays the error graph for regression imputation method. It can be observed from Figure 4.21 that interpolation has performed poorly with long consecutive missing gaps and the error difference between the predicted value and real values at point is near 10 which can really affect the analysis phase. The error graph for regression has shown better performance as compared to the interpolation imputation method. The error graph

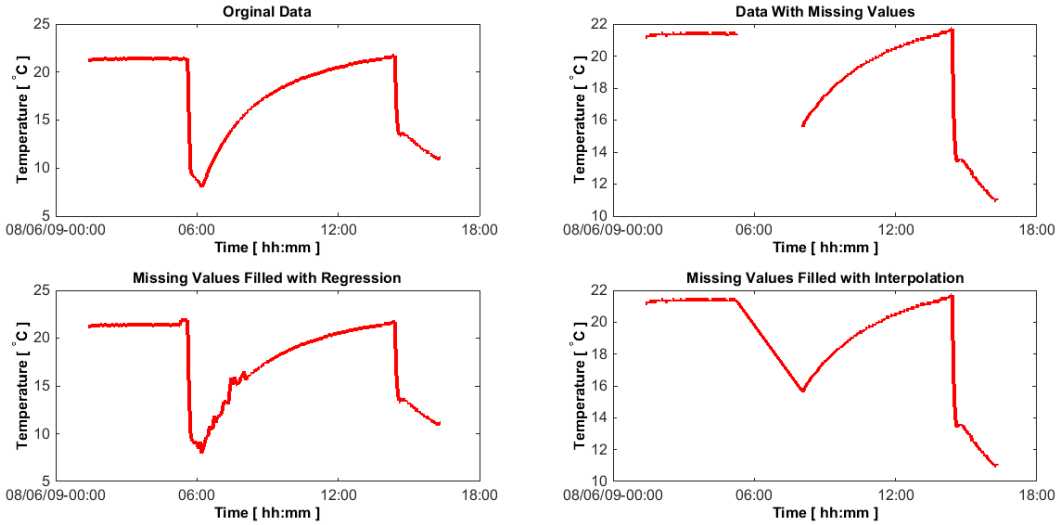


Figure 4.20: Filling longer gaps of missing data with linear interpolation and regression

for interpolation i.e. the first graph in Figure 4.21 has error values far away from 0 showing more difference in places where predictions are made. As compared the error graph of regression shows error values near to 0, meaning less difference in actual and predicted value.

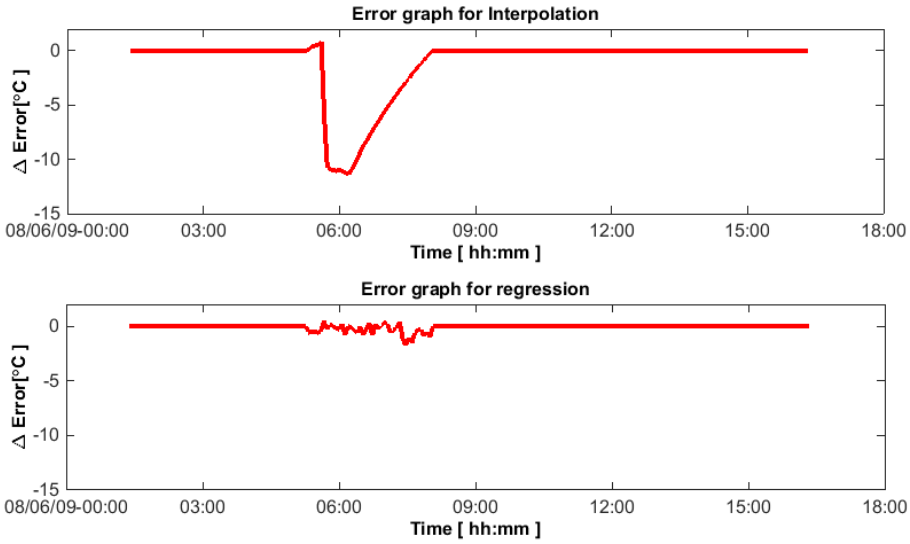


Figure 4.21: Error graph for longer gaps with linear interpolation and regression

4.4.1 Proposed Data Imputation Method

As data imputation is an important step, as it can greatly influence the analysis phase of the data, especially for finding various patterns in the operational data of energy systems in buildings in order to detect and diagnose faults. In order to handle the missing gaps, a data imputation method using interpolation has been proposed in this section. As discussed in the previous section that the interpolation method has produced better results as compared to interpolation, therefore, the proposed method uses interpolation for data imputation. The larger gaps are ignored as the

reason is that filling larger gaps will not benefit in the analysis phase. And at the same time filling the larger gaps will not add any additional patterns or information, and there is high possibility that decisions made on these results are not reliable. The main issue with predicting the large gaps is the randomness in the parameters such as weather parameters etc. Therefore the proposed method interpolates data for shorter period of time and neglects data imputation for larger gaps. The Algorithm 4.8 takes the $Time\_Stamp$, $Data$, $Required\_time\_interval$, Num_Ignore_NaN as input and creates the new time stamp data with the required interval.

Algorithm 4.8: Data_Imputation Pseudo Code

```
1  Input: Time_Stamp, Data, Required_time_interval, Num_Ignore_NaN;
2  begin
3  interval = Find_interval(Time_Stamp, Data)
4  data_mark = Data_Marker(Time_Stamp, Data, Num_Ignore_NaN)
5  int_test = Check_integer(Data)
6  for i = 1 : Max(data_mark) do
7      temp_data := find(data_mark = i)
8      temp_timestamp := find(data_mark = i)
9      [hold_newtime hold_newdata] :=
       Data_Resample(temp_timestamp, temp_data, Required_time_interval, int_test)
10     New_Time_Stamp[i] := hold_newtime
11     New_Data[i] := hold_newdata
12 end
13 return New_Time_Stamp, New_Data
14 end
```

There are different functions used in Algorithm 4.8, while handling missing gaps in data of energy system of the building. Each function is addressing an issues faced during the implementation of the algorithm. The following are the description of the function (issues) that were handled during the data imputation,

- The first step is to automatically detect whether the given data is integer or real. The reason behind this step is when we are creating new time stamps for data integer type data such system state which is either 0 or 1, the interpolation should return the data in integer format. Algorithm 4.9 is used for this purpose. The input to the function is the data for which the data type is required, whereas, the function returns binary value representing

Algorithm 4.9: Check_integer Pseudo Code

```
1  Input: Data;
2  begin
3  Initialize: output := 0
4  if round(Data) − Data = 0 then
5      output := 1
6  end
7  return output
8  end
```

whether the data is integer or not.

- In order to be able to interpolate only small gaps under certain threshold, Algorithm 4.10 is used to mark the areas for interpolation and ignore if the missing gaps are larger. The parameters passed to the algorithm are data, time stamps of the data, and the number of consecutive missing gaps to be considered for interpolation. The gaps that are larger are marked so that interpolation for that period is not used and help in analysis phase, as this will help to identify areas where data is not available. The function returns the $Mark$ parameters showing the gaps to be filled with interpolation and areas where existed the longer gaps in data.

Algorithm 4.10: Data_Marker Pseudo Code

```
Input: Time_Stamp, Data, Num_Ignore_NaN;
begin
Initialize: count :=  1, Mark := 0
if length_of_Data > Num_Ignore_NaN then
    for i = 1 : length_of_data do
        Mark[i] := count
        if isnan(Data[i : i + Num_Ignore_NaN]) then
            if i > 1 then
                count := count + 1
            end
            Mark[i : i + Num_Ignore_NaN] := count
            count := count + 1
        end
    end
end
return Mark
end
```

- One of the problems that have been faced was the data having different time stamps and different intervals between the two consecutive time stamps. In case of the different time intervals, it will create more difficulty, as analyzing some features of the energy system at one time stamp will not be possible. Therefore, it is required to automatically detect the interval for each data point. In order to find the interval between consecutive data points

Algorithm 4.11: Find_interval Pseudo Code

```
Input: Time_Stamp, Data;
begin
Initialize: time_difference := 0
for i = 2 : size_of_data do
    time_difference := Time_Stamp[i] − Time_Stamp[i − 1]
end
interval:= mode(time_difference)
return interval
end
```

automatically Algorithm 4.11 is used. The algorithm considers the most frequent interval as the interval for the whole data point. The input to the function is the time stamps of the data, and data for which interval has to be found out. Once the interval of the data is known, the data can be interpolated according to the required time stamp, and all parameters can be available for a specific time stamp.

- In this step the algorithm generates the new time stamp according to the desired interval and generates the data using the interpolation method while ignoring the large gaps as found out in the previous step. Algorithm 4.12 is used for re sampling of data as per required interval. The input to the algorithm were the time stamp of the data, data values, the desired time interval for data, and the type of data as integer or real. The function return the time stamps with desired interval along with the data values interpolated.

Algorithm 4.12: Data_Resample Pseudo Code

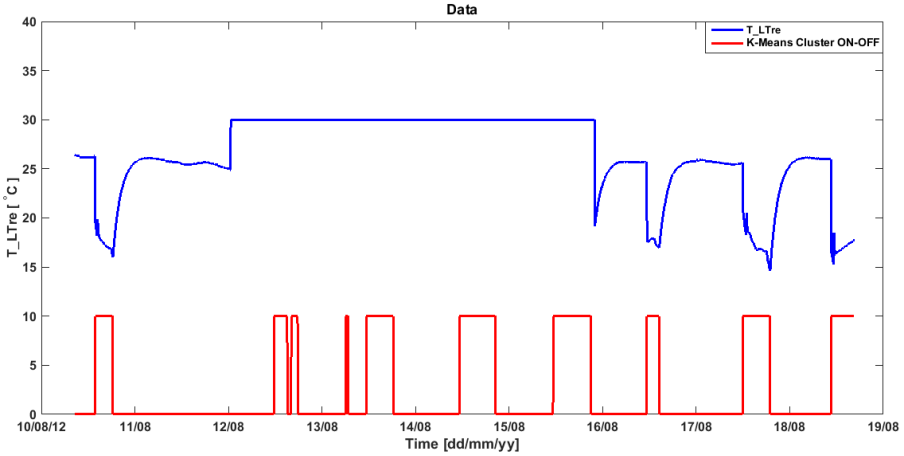
```
1  Input: Time_Stamp, Data, Required_time_interval, check_integer;
2  begin
3  Initialize: Start_Time := round(Time_Stamp[1])
4  Initialize: End_Time := round(Time_Stamp[end])
5  New_Time_Stamp := Start_Time : Required_time_interval : End_Time
6  New_Data := resample(Time_Stamp, Data, New_Time_Stamp)
7  if check_integer = 1 then
8      New_Data := round(New_Data)
9  end
10 return New_Time_Stamp, New_Data
11 end
```

5 Anomalies Detection

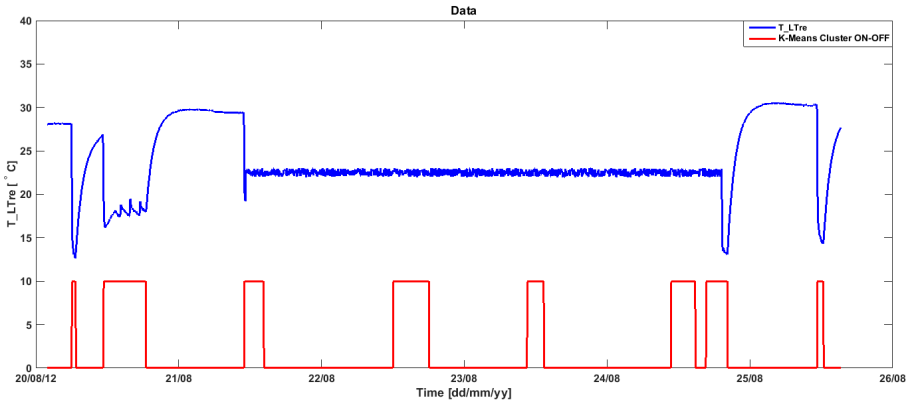
The recent advancement in sensor technology has enabled to record the real life phenomenon like temperature, pressure etc. of buildings for longer period of time. The sensors in the field are generally operating in the non-controlled environment; therefore the sensor fault probability in data is very high. At the same time, the large amount of data recorded in energy efficient buildings has made it a challenging task to detect such faults manually. In this chapter, the proposed method for automatically detection of such fault patterns in sensor data of water chillers is discussed. The input of the proposed technique is the operational state information of the chiller obtained by using the k-means algorithm. The data is transformed to a symbolic representation by applying the Symbolic Aggregate Approximation (SAX) method. The SAX symbols are converted to Bag of Words Representation (BoWR) for each On/Off cycle, which allows for automated detection of a class of sensor faults. The method is applied to the real life data of a water chiller. The results shows improvement in comparison with state of the art techniques.

There are different types of faults patterns found during the analysis of the data recorded by sensors in faulty state. The different faulty pattern can be either a Short faults (singleton outliers or abrupt change in the successive data), Noise faults (the variance in the sensor data is increased), or a Constant faults (the value of the sensor is stuck at one value) that causes the readings of the faulty data to deviate from the normal pattern demonstrated by a real sensor reading. It is important to understand these sensor fault patterns and develop methods for automatically detecting them in the data.

The method for automatic detection of singleton outliers (Short) has been discussed in Section 4.3 of the thesis. The focus in this chapter is to find a method for the automatic detection of observed Constant faults and Noise fault patterns in the real life data of chiller as can be seen in Figure 5.1. The Constant fault pattern can be observed in Figure 5.1(a), whereas, Figure 5.1(b) shows the Noise fault pattern. The basic concept in the proposed technique called as bag of words representation (BoWR), is to use the duty cycle (On/Off) information for the detection of sensors faults. As the behavior of the chiller varies in these two states, therefore, the patterns of the consecutive cycles are compared for similarity. If the patterns of the consecutive cycles are similar, then it will be considered as a sensor fault. The proposed BoWR method has been applied to a real life data of energy systems in builind such as chiller, and is compared with the histogram method from [54, 59], autoregressive integrated moving average (ARIMA) model [54] and artificial neural network (ANN) [162] based methods.



(a) CONSTANT fault pattern



(b) NOISE fault pattern

Figure 5.1: Sensor fault patterns in data

5.1 Methodology

In this section the methodology adopted for finding the patterns of anomalies is discussed. In first case the duty cycle(On/Off) has been detected using the k-means clustering algorithm which is discussed in the 4.1.2.2 section of the thesis. In the next phase the data is normalized using z-score, followed by discretization process using symbolic aggregate approximation. Subsequently, a bag of word representation (BoWR)of each duty cycle (On/Off) is created. For finding the anomalies in the sensor data, the bag of word representation pattern of consecutive duty cycles are compared. As it is expected that energy systems in building perform differently in operational and non-operational cycles, therefore in case of having same patterns in consecutive cycle is an indicator of anomaly in sensors data.

5.1.1 Symbolic Aggregate Approximation (SAX) Transformation

The first step performed for this process is the detection of duty cycles (On/Off) cycles. After detection of duty cycles, the data will in the form as defined by the Equation 5.1,

$$C_i = \{S_1, S_2, S_3, ..., S_N\}, \tag{5.1}$$

Where C_i is the i^{th} cycle and S_t is the sensor value at time tick k. The data is normalized using Z-Score normalization which is given as,

$$Z(Data) = \frac{S_t - \mu}{\sigma}, \tag{5.2}$$

Here $Z(Data)$ is the z-score normalized form of the data, and S_t is the sensor data at t^{th} time tick, μ represents the mean and σ is the standard deviation of the whole data. After applying the z-score normalization, the data will be now in the form as defined in Equation 5.3 while using the On/Off information from the k-means clustering algorithm,

$$Cycle\_i = \{Z_1, Z_2, Z_3, ...Z_{N_i}\}, \tag{5.3}$$

Where $Cycle\_i$ is the i^{th} cycle of the data and in case if odd count of i is representing Off cycle then even count of i will be presenting the On cycle in data. Z_t is the normalized sensor value while N_i is the event count of $Cycle_i$.

Each cycle data is first broken down into M non overlapping sub-sequences, in a uniform manner, just like the example illustrated in Figure 5.2, wherein the partitions are represented by alphabets a, b, c. This process is called as chunking, and the period (x-axis) can be of different time length (P) depending on the application where it is used [12]. The value of P is taken as 5 minutes in this research. The symbol of each data point is assigned according the breakpoints. The number of break points (M) taken for this research is 60.This transforms the data for each cycle to symbols. The SAX representation is specific for each a length of each cycle. In order to generalize the symbolic representation for each cycle with different lengths, the BoWR is used.

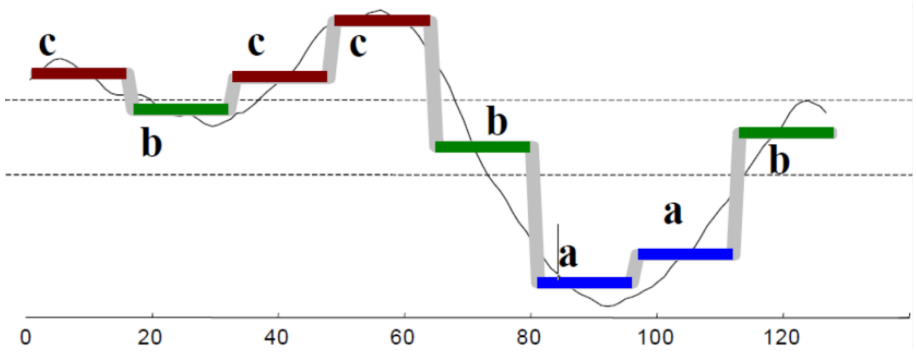


Figure 5.2: Example of Symbolic Aggregate Approximation (SAX) transformation of On/Off Cycle with M=3 and P=20 [163]

5.1.2 Bag of Words Representation (BoWR)

The SAX transformation will transform the time series data in the symbols. The value of M is taken as 60 for this research. The bag of words representation can be defined by a vocabulary set $\{w_1, w_2, w_3, ..., w_M\}$ and the associated histogram vector $BoWR_i$ for i^{th} cycle:

$$\mathrm{BoWR_i} = \left(\mathrm{V_1^i\ V_2^i\ V_3^i\ .....\ V_M^i}\right), \tag{5.4}$$

where V_j^i is the number of occurrences of w_j in the i^{th} cycle, i.e.

$$\mathrm{V_j^i = Count_i(w_j)}, \tag{5.5}$$

here the subscript i in $Count_i$ refers to the i^{th} cycle.

In order to handle the different time length cycles a better idea is to normalize, i.e. use relative frequencies. With this in view, Equation 5.5 has been modified as Equation 5.6 given below, where N_i is representing the number of time ticks in the i^{th} cycle

$$\mathrm{V_j^i = \frac{Count_i(w_j)}{N_i}}. \tag{5.6}$$

The Table 5.1 shows the example of how the bag of words representation will denote each cycle. The first column shows the cycle number, whereas, the next columns represent the count number of each word in the cycle.The benefit of using BoWR method is that it will reduce the dimensions of data as well.

Table 5.1: Example of SAX Vocabulary table

SAX Vocabulary										
Data	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	...
BoWR ($Cycle_1$)	0	5	4	3	2	1	0	0	1	...
BoWR ($Cycle_2$)	1	9	7	4	5	3	1	5	2	...
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
BoWR ($Cycle_{last}$)	...	...	...	...	...	...	...	...	...	...

5.1.3 Algorithm for Sensor Fault Detection

The Algorithm 5.1 presents the procedure for finding the sensor faults. The On/Off cycle information that has been automatically detected using k-means can be used in the algorithm. The basic idea of the algorithm is to use the On/Off state information and compare the patterns of the consecutive cycles (which will be On cycle followed by a Off cycle or vice versa). The expectation is that the machine will generate different patterns (behavior) in the two states, during the normal working of the machine.

The input to the algorithm is the bag of word representation (BoWR) of all the cycles and the threshold (δ) which is the value of similarity required to consider the consecutive cycle patterns as sensor fault. In case of *Constant* sensor fault the value will be equal to zero while for $Noise$ fault patterns it will be very low.The algorithm returns a vector $fault$, having a size equal to the total number of cycles which are passed to it. After the execution of the algorithm, the value $fault$ at subscript i, i.e. $fault[i]$ would be 1 (true) if fault detected at cycle i, otherwise it would be 0 (false). The *dist* function in the Algorithm 5.1 is used for finding the difference between two cycles. In this research the Manhattan distance function is used for finding the similarity and M is taken as 60, while the value for the δ is chosen as 1. The algorithm is tested with the data of absorption chiller and results are compared with the state of the art methods.

Algorithm 5.1: Sensor fault detection

```
Input: BoWR[num_of_cyles, M], δ;
begin
initialize fault:= zeros[num_of_cyles];
for i = 1 : num_of_cyles − 1 do
    if dist (BoWR[i, M], BoWR[i + 1, M]) < δ then
        fault[i] = 1;
        fault[i + 1] = 1;
    end
end
return fault;
end
```

The steps involved for the detection of sensor faults can be summarized as:

- The On/Off cycle information of the energy system will be required for the detection of anomalies algorithm. The duty cycle (On/Off) information is automatically detected by using the k-means clustering algorithm.
- The data is normalized by using the z-score normalization.
- The data of each cycle is transformed to symbols by using the symbolic aggregate approximation (SAX) transformation method with vertical breakpoints (M) equal to 60 and horizontal breakpoints (P) equal to 5 minutes (There time interval between two consecutive time stamps is 5 minutes for the data of the system that has been taken under consideration).
- The bag of words representation (BoWR) has been created for each cycle of the data.
- Use the information of On/Off cycle detected by using k-means algorithm and compare the BoWR of consecutive cycles. If the BoWR of consecutive cycles is similar then mark those cycles as anomalies.

5.2 Experiments and Results

This section discusses the experimentation and their results from the methodology adopted for detecting the anomalies patterns. These anomalies are found due to various reasons such as noise in the communication link or sensor fault etc. It is important to detect these kind of anomalies as this will help to take these areas under consideration during the analysis phase.

The data from various energy systems such as water based chillers has been used for different experiments for detecting the anomalies patterns in the data. The proposed BoWR method has been applied to a real life data of chiller. The proposed method is compared with the already existing state of the art methods such as histogram method from [54, 59], autoregressive integrated moving average (ARIMA) model [54] and artificial neural network (ANN) [162] based methods.

The first step for bag of words representation (BoWR)is to normalize data with z-score. The next step is to transform the z-score normalized data in to Symbolic Aggregate Approximation

(SAX) symbols. Figure 5.3 shows the SAX transformation (dashed line) along with the Z-Score representation of data (T_LTre). The number of breaks M is taken as 60, that makes the approximation by SAX representation quite near to the real data, whereas, the value of P is considered as 5 minutes.

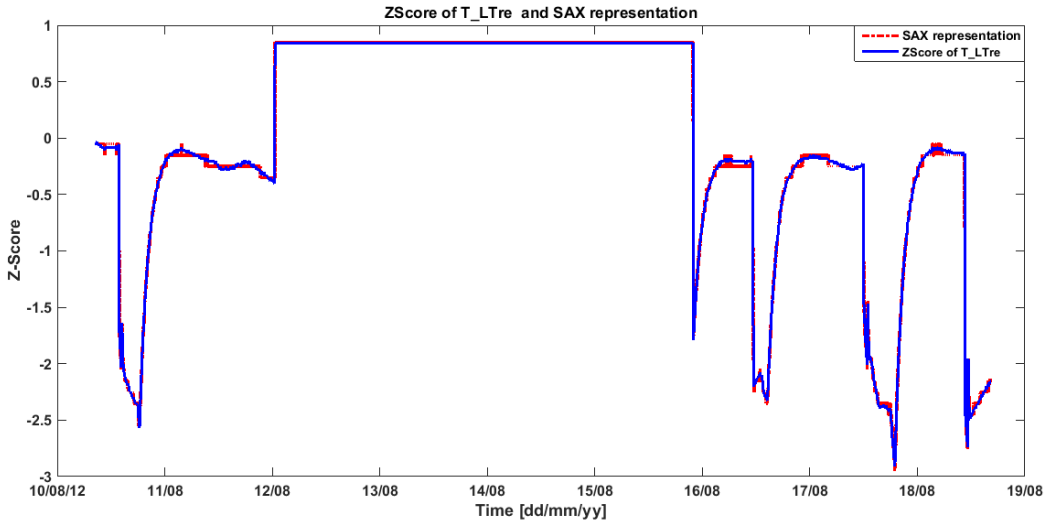
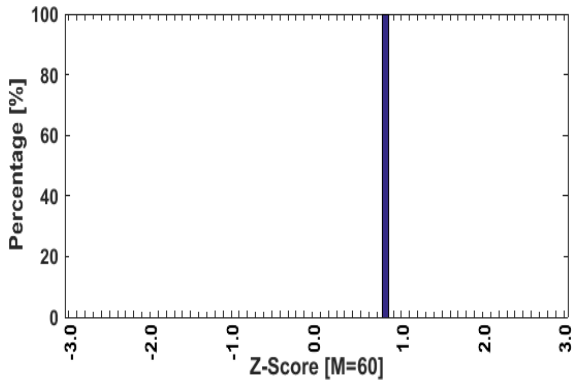
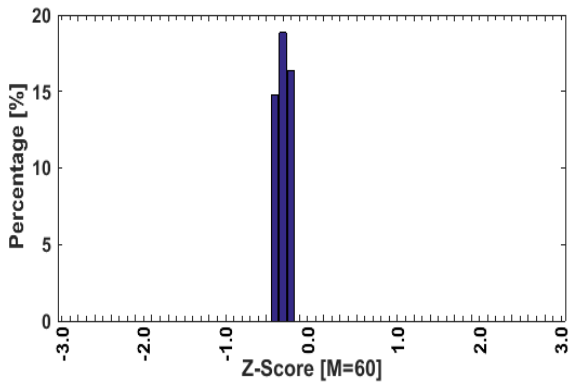


Figure 5.3: Z-Score representation and Symbolic Aggregation Approximation (SAX) representation of T_LTre

The duty cycle (On/Off) information has been detected using k-means clustering algorithm that has been discussed in Section 4.1.2.2 of the thesis. In the next phase the SAX symbols of each cycle are transformed to bag of word representation (BoWR). The data is in the form of z-score as can be seen in Figure 5.4(a), where the BoWR representing the *Constant* value in cycle has been shown with histogram. The x-axis values are the symbols representing the 60 breakpoints of values $\mu \pm 3\sigma$, since BoWR entries are normalized via z-scores and divided in ($M = 60$) breakpoints. Similarly the Figure 5.4(b) is representing the *Noise* pattern histogram, where the data is distributed to only few blocks.



(a) Bag of Word Representation histogram of one cycle during *Constant* sensor fault



(b) Bag of Word Representation histogram of one cycle during *Noise* sensor fault

Figure 5.4: Histogram representation of data

The bag of word represention (BoWR) of 120 consecutive cycles are shown in Figure 5.5 and Figure 5.6, in order to get the view of the patterns of *Constant* fault and *Noise* faults in data The detection of *Constant* sensor fault can be observed from the graphs given in Figure 5.5 wherein BoWR is shown against each of the 120 cycles. The graph visualize the data density in each BoWR as a color intensity map given on the right side of the graph. The abscissa values are the 60 break points of $\mu \pm 3\sigma$, since BoWR entries are normalized via z-scores. Each horizontal pattern of colors in Figure 5.5 represent a BoWR of the corresponding cycle. If one peak in observed in the graph, then it's definitely a sensor fault showing a constant pattern. In Figure 5.5, the *Constant* fault patterns can be seen in the data where the BoWRs for cycle $BoWR[8]$ through $BoWR[56]$ exhibit exactly one frequency that has been represented with the highest intensity (brown color in the intensity map). The dash line rectangle with arrow is denoting the area for *Constant* fault in the Figure 5.5 . The absence of fault is demonstrated by dilated BoWR entries, e.g. $BoWR[57]$ to $BoWR[120]$ in Figure 5.5 have no fault.

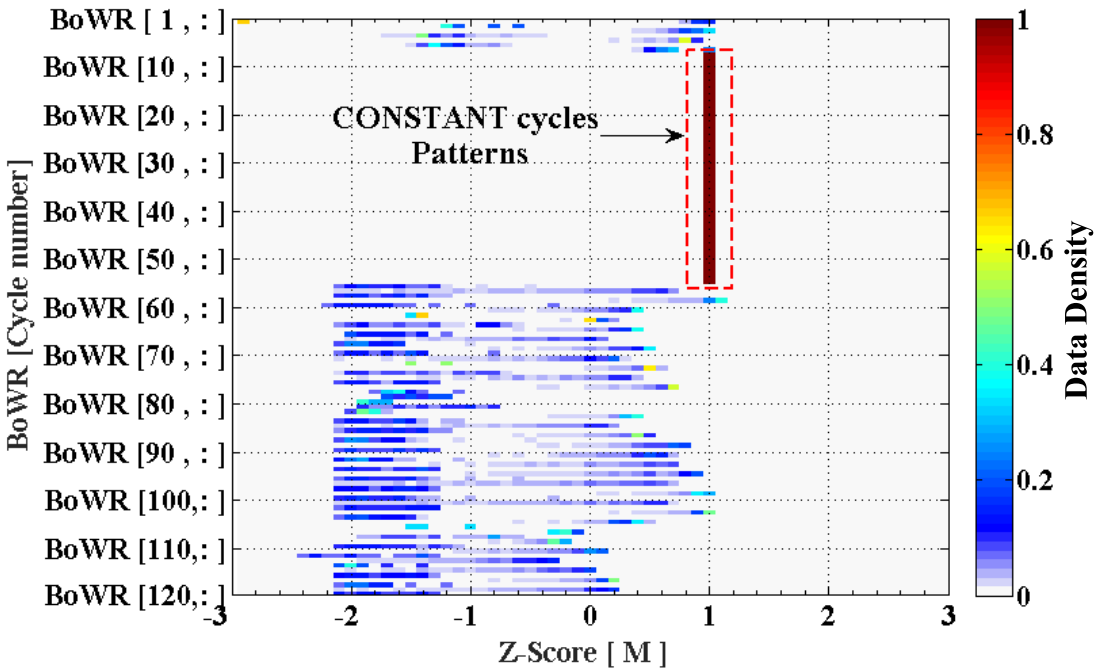


Figure 5.5: CONSTANT fault pattern

The detection of *Noise* sensor fault can be visualized from the graph given in Figure 5.6 wherein BoWR is shown against each of the 120 cycles. The x-axis values are the 60 break points of $\mu \pm 3\sigma$, since BoWR entries are normalized via z-scores. Each horizontal pattern of colors is representing the BoWR that can be observed in Figure 5.6.It has been observed that being a relative frequency histogram, the consecutive BoWR have significant 'spread' in the absence of any fault. In other words, if BoWR has fewer frequencies (lower 'spread) and leptokurtic in consecutive cycles then it points to some fault. It is quite possible that faulty BoWR may exhibit more than one peak due to the presence of sensor noise in the consecutive cycles. The *Noise* sensor faults patterns are demonstrated in Figure 5.6, wherein $BoWR[32]$ to $BoWR[104]$ have similar BoWR patterns, restricted to just a few frequencies, as compared to others cycles; thus marked as *Noise* sensor fault. The dash line rectangle with arrow is denoting the area for *Noise* fault in the Figure 5.6. The absence of fault is demonstrated by dilated BoWR entries, e.g. $BoWR[1]$ to $BoWR[30]$ and $BoWR[106]$ to $BoWR[120]$ in Figure 5.6 shows normal behavior having no fault.

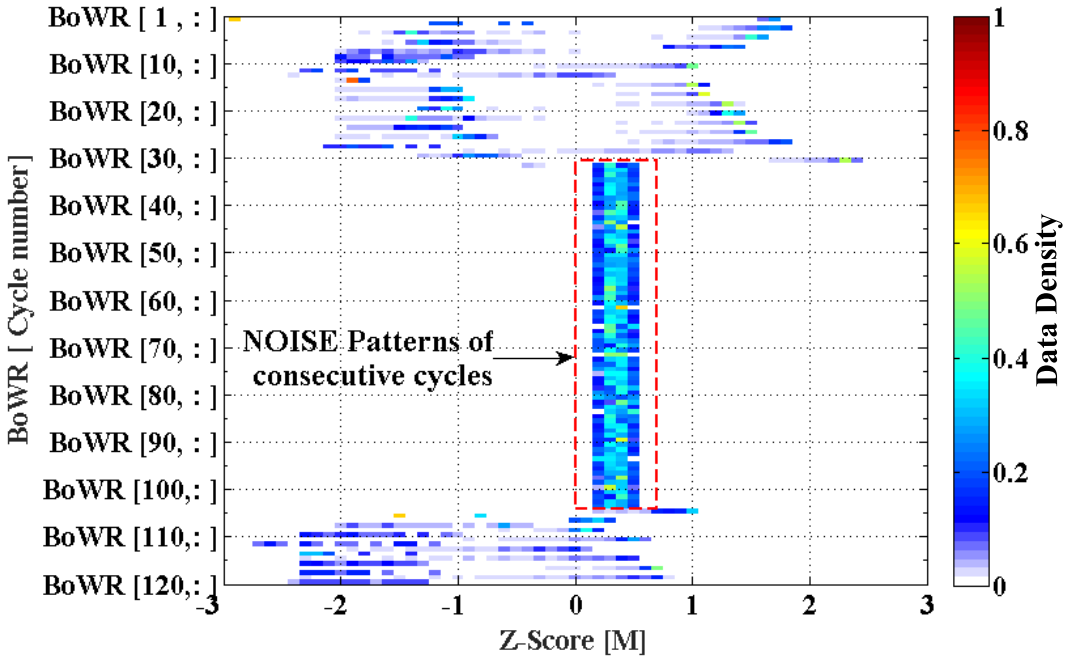


Figure 5.6: NOISE fault pattern

5.2.1 Comparison With Reference Methods for Sensor Fault Detection

The different tests are performed for checking the accuracy and performance of histogram method, autoregressive integrated moving average (ARIMA), neural network (ANN) and the proposed bag of word representation (BoWR) method. The comparison of the methods is done on the basis of parameters discussed below. The required information for parameters is explained in Figure 5.7, where the columns represent the number of events found in the real data and rows represent the predicted events by the method.

Observed Events

Predicted Events	Yes	No
Yes	True Positives	False Positives
No	False Negatives	True Negatives

Figure 5.7: Contingency table

- **Probability of Detection**: It indicates the number of faults occurred in the data, being detected by the used method [164], which can be calculated using,

$$\text{Probability of Detection} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5.7}$$

- **False Alarm Rate**:It is the number of times that an event was incorrectly predicted [164], given as,

$$\text{False Alarm Rate} = \frac{\text{False Positives}}{\text{True Positives} + \text{False Positives}} \tag{5.8}$$

- **Accuracy**: It indicates the overall correct predictions [165, 166], can be calculated as,

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{5.9}$$

The histogram algorithm method was employed with different configurations each depending on the fact that whilst histogram method is dependent on window size (N). The results can be seen in Table 5.2 for *Constant* and *Noise* sensor faults. It has been observed that histogram method has performed better with *Constant* faults, particularly with small window size (N) such as N equal to 10, where the highest *Probability of detection* and *Accuracy* for Constant faults is achieved with 99.83% and 92.40% respectively, along with lowest *False Alarm Rate* of 18.09%. However the histogram method has not performed the best for *Noise* faults. The *accuracy* with smaller window size (N) has shown 50.70%, which decreases with the increase in the window size(N). Moreover, the higher values of N has shown little improvement in the *Probabilty of detection* and *False alarm rate*.

Table 5.2: Histogram results with different window size(N) values

Methods	**CONSTANT FAULTS**			**NOISE FAULTS**		
	Probability of Detection (%)	False Alarm Rate (%)	Accuracy (%)	Probability of Detection (%)	False Alarm Rate (%)	Accuracy (%)
Histogram (N=10)	**99.83**	**18.09**	**92.40**	0.02	99.86	**50.70**
Histogram (N=100)	99.51	36.90	79.91	0.65	97.83	46.03
Histogram (N=500)	99.51	45.84	71.00	1.90	96.96	33.45
Histogram (N=1000)	91.85	50.01	65.76	**1.90**	**96.96**	33.45

The autoregressive integrated moving average (ARIMA) method was tested with different configurations each depending on the fact that ARIMA method is dependent on look ahead steps (L). The results for ARIMA can be seen in Table 5.3 for *Constant* and *Noise* sensor faults. It was noteworthy to see that ARIMA method has performed better with *Noise* fault, specially with larger look ahead step(L) such as L equal to 1000, where the highest *Probability of detection* and lowest *False Alarm Rate* of 88.01 and 83.62 respectively were achieved. The best *Accuracy* for *Noise* faults was achieved with L=10 having 77.88%. However the ARIMA method has not performed the best for *Constant* faults. The *accuracy* with smaller look ahead step (L)=10 has shown 64.07%, which decreases with the increase in the look ahead steps(L). Moreover, the higher values of L has shown little improvement in the *Probabilty of detection* and *False alarm rate* for *Constant* faults.

Table 5.3: ARIMA model results with different look-ahead (L) values

Methods	**CONSTANT FAULTS**			**NOISE FAULTS**		
	Probability of Detection (%)	False Alarm Rate (%)	Accuracy (%)	Probability of Detection (%)	False Alarm Rate (%)	Accuracy (%)
ARIMA (L=10)	2.66	98.79	**64.07**	9.29	96.84	**77.88**
ARIMA (L=50)	14.94	95.87	50.80	12.59	96.64	73.51
ARIMA (L=100)	17.98	95.19	50.18	12.75	97.09	63.35
ARIMA (L=1000)	**48.62**	**91.34**	35.73	**88.01**	**83.62**	72.83

The comparison of the various state of the art methods (histogram, ARIMA and artificial neural network (ANN)) with propose bag of word representation (BoWR) method. The results can be seen in Table 5.4 for *Constant* and *Noise* sensor faults. It has been observed that histogram method has the best *Probability of detection* with 99.83% for *Constant* faults. Moreover, artificial neural networks (ANN) has achieved the least *False Alarm Rate* with 0.10%, while the highest *Accuracy* has been achieved by BoWR with 98.77% for *Constant* faults. The BoWR dominates in all the comparison parameters for detection of *Noise* faults. The BoWR has achieved the highest *Probabilty of detection* with 99.84%, lowest *False alarm rate* of 1.16% and best *accuracy* with 99.44% for detection for *Noise* faults.

Table 5.4: Sensor fault detection performance of different methods.

Methods	**CONSTANT**			**NOISE**		
	Probability of Detection (%)	False Alarm Rate (%)	Accuracy (%)	Probability of Detection (%)	False Alarm Rate (%)	Accuracy (%)
Histogram (N=10)	**99.83**	18.09	92.40	0.02	99.86	50.70
Histogram (N=1000)	91.85	50.01	65.76	1.90	96.96	33.45
Arima (L=10)	2.66	98.79	64.07	9.29	96.84	77.88
Arima (L=1000)	48.62	91.34	35.73	88.01	83.62	72.83
Neural Network	96.00	**0.10**	97.20	96.34	3.40	95.90
BoWR	97.75	1.39	**98.77**	**99.84**	**1.16**	**99.44**

Figure 5.8 shows the stuck value pattern detection. The Figure 5.8 has two graphs. The first graph is representing the real data along with the detection line at the top (red color of line shows where stuck at value is detected, whereas green color of the line is representing no anomaly detected). The last graph is showing the duty cycle detected with k-means clustering algorithm. The dotted line rectangle is highlight the existence of the stuck at value anomaly as can be noticed in Figure 5.8.

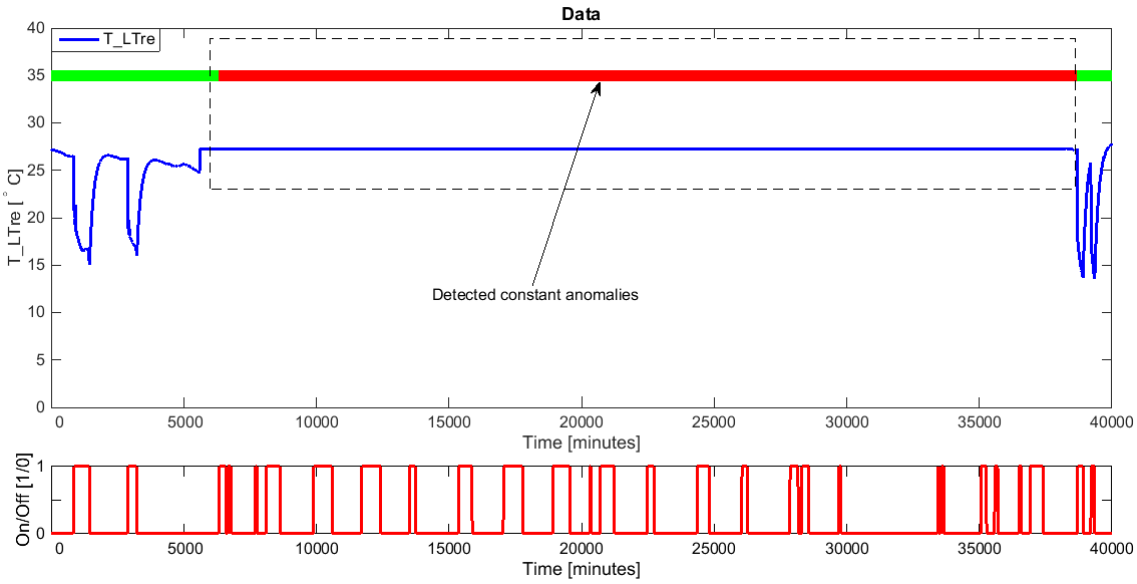


Figure 5.8: Detection of Sensor faults (Stuck at value)

Figure 5.9 shows the noise pattern detection. The Figure 5.9 shows two graphs the same as in the previous figure. The dotted line rectangle is highlight the existence of the noise anomaly as can be noticed in Figure 5.9.

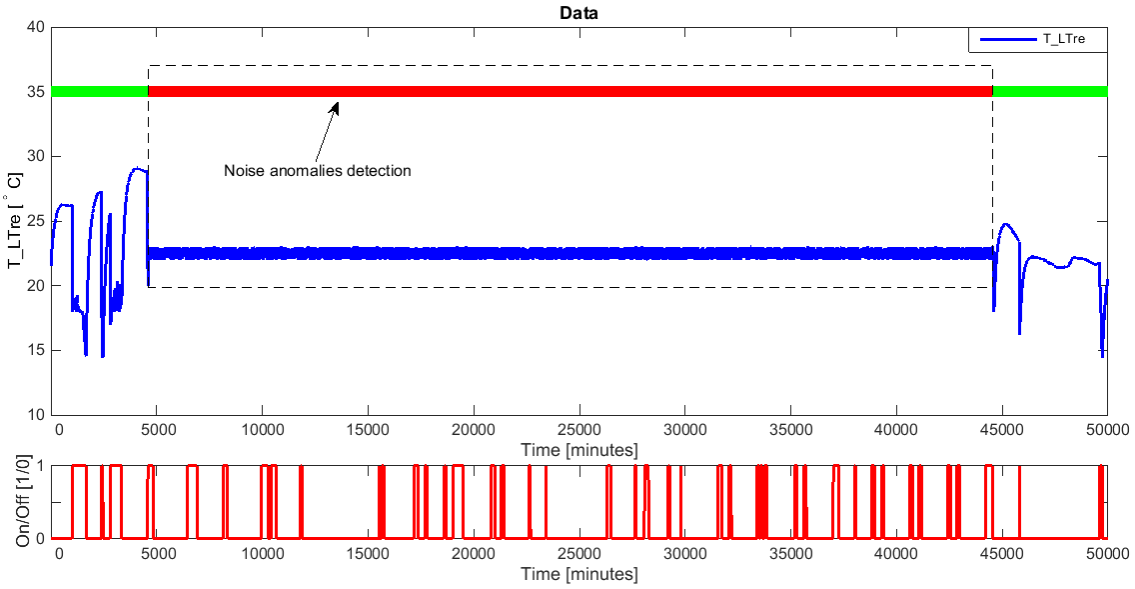


Figure 5.9: Detection of sensor Faults (Noise)

6 Fault Patterns Detection and Diagnosis in Operational Behavior of Energy Systems

The buildings are monitored using sensors in order to capture the different complex energy parameters for analysis. As a result, a lot of raw data is recorded during the monitoring of these buildings. The easiest way adopted by the experts in the field is to use a variety of visualization tools for the analyses of the data. But, the huge amount of data recorded during monitoring makes it difficult for the experts to have a detailed performance analysis of the buildings manually or with simple visualizations. Moreover, there is a high chance for overseeing some important patterns in the data that may lead to faults in the different components of building, causing reduction in the energy efficiency.

The methods that can automatically extract the different patterns in buildings data, may help the experts to get more insight to the various parameters of the buildings energy usage as well as other processes, like the source of faults in different components of the buildings. Data mining techniques, like clustering can help in automatically detecting the different patterns in buildings data [12, 13, 14]. The process of automatically finding various patterns in the data can make the subsequent analysis easier, feasible and lesser laborious [12, 13, 15].

In order to automatically find different patterns in the absorption chiller's operation, in this chapter, we propose to use a bag of words representation (BoWR) with subsequent hierarchical clustering. The proposed method has been applied to a real life data of absorption chiller and compared to another approach called dynamic time warping (DTW). The On/Off state information required for the suggested technique was taken from the k-means clustering algorithm as discussed in Section 4.1.2.2. As we are taking the sensor readings that are placed outside the chiller, therefore the sensors reading will reflect the behavior of the chiller during the operational cycle of the chiller. Thus, the On (operational) cycles are of greater importance for finding the performance of chillers and faults detection and diagnosis (FDD). Moreover, avoiding the Off cycle for finding different patterns will reduce the amount of data as well. The same procedure will be used for BoWR transformation used for clustering as discussed in Section 5.1.2. At the end, the results of hierarchical clustering with BoWR representation and DTW method are compared using cophenetic correlation.

After clustering the data, additional information is added to each cluster such as average CoP of the cluster and average time of On cycles in a cluster. This information will be used for diagnosis of faults in the chiller. In order to be able to diagnosis various faults in the chiller, a rule based classification algorithm has been suggested and discussed. Furthermore, additional parameters

of operation have been integrated into the analysis, enabling to diagnose these the faults. The concepts discussed in this chapter are published in [167].

6.1 Detection of Fault Patterns

In this section the methodology adopted for the detection of various patterns in the energy system, specifically the faulty patterns has been discussed. The analysis of air ventilation system has been discussed followed by the section analysis of the chiller operational data. A bag of words representation (BoWR) method, subsequent with hierarchical clustering has been adopted for the detection of faulty patterns in data.

6.1.1 Analysis of Air Ventilation System Operational Data

The data from a ventilation system has been taken under consideration and analyzed. In order to see the operational data behavior of the ventilation system, heatmaps of the supply air pressure an exhaust air pressure has been drawn as can be seen in Figure 6.1. The y-axis in Figure 6.1 is representing the days of the week, whereas, the x-axis is denoting the hours of the day starting from 0 to 24 hours. The values of the pressure are represented with color (color values are represented with colorbar attached at right side of each figure) against time and day of the week. The white color is representing 0 pressure, wnereas, the red color represents the maximum pressure. The behavior of the ventilation system looks plausible from Figure 6.1 as the ventilation system works from morning around 7:00 a.m. to night time around 9:00 p.m. And on Saturday the ventilation system is running till around 1:00 p.m. The reason is, data have been recorded from a seminar room, where there are lectures on saturday till 12:00 o'clock. On Sunday the ventilation system is off as there are no occupants.

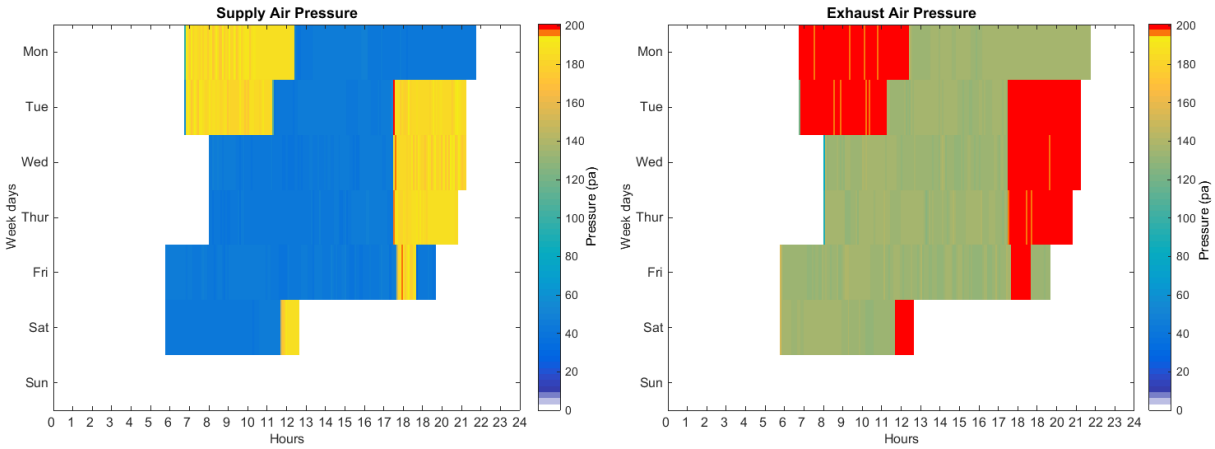


Figure 6.1: Ventilation Air Pressure heatmap

The same kind of behavior can be extracted from Figure 6.2, where the data of supply volume flow and exhaust volume flow are displayed against the days of the week and the hour of each day. The volume flow parameters also show the same pattern of air ventilation system behavior of each day as in the case of air pressure.

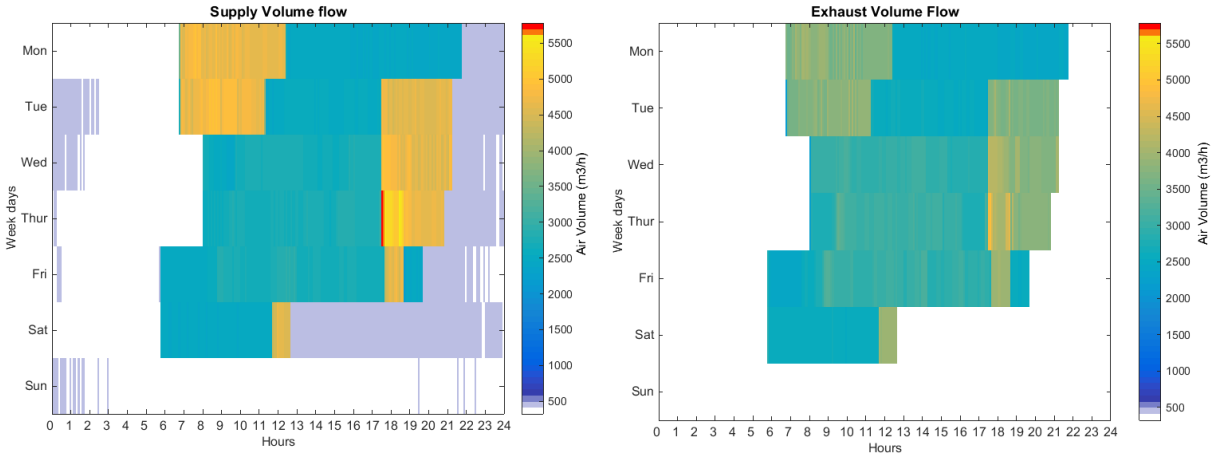


Figure 6.2: Ventilation Air Volume Flow heatmap

The Figure 6.3 represents the VAV signal for the supply and exhaust ventilation air volume (VAV) against the days of the week and the hour of each day. This signal basically shows the air control of the building. The supply VAV signal shows the percentage of air supplied to the building, whereas, the VAV signal exhaust shows the air percentage, that has been taken out of the building. The values of the VAV signal are represented in percentage and ranges between 0% to 100% as can be observed in Figure 6.3. The VAV signals show that the air ventilation was working on week days and Saturday till around 14:00 hour, and was not operational on the Sunday. The operation of the VAV signal shows the acceptable behavior of the ventilation system.

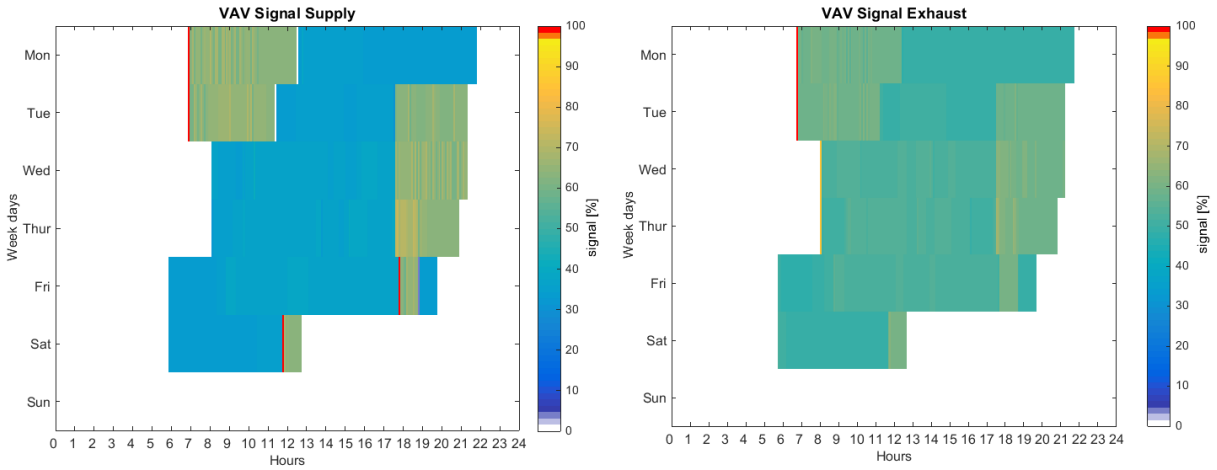


Figure 6.3: Ventilation Air Volume Signal for building heatmap

The k-means clustering algorithm is also used for finding the presence of the occupants in the building, using the air quality information (temperature, relative humidity and carbon dioxide) of a ventilation system inside the building. The ventilation system that is analyzed uses the different sensors for calculating the temperature, humidity and carbon dioxide in the building. These values are used to handle the ventilation systems for maintaining good air quality in the building. The automatic detection of occupants can help in efficiently managing the ventilation system and lightening system in the building as well, which will help in saving energy. The Figure 6.4

shows the behavior of the sensors of air quality inside the building. The x-axis in Figure 6.4 is representing the time, and it can seen that the indoor environment in the building is maintained as the temperature range is between $21°C$ and $24°C$, the relative humidity is kept around 38% during the presence of occupants in the building, and carbon dioxide level below $1500ppm$. The first graph in the Figure 6.4 shows the temperature sensors placed at different locations in the building, while the second graph is representing the relative humidity in the building, and the third graph displays the carbon dioxide level maintained in the building. The fourth graph is representing the variable air volume (VAV) signals showing the air flow with value between 0% to 100%.The last graph in the Figure 6.4 shows the k-means results for finding automatically the occupants in the building. When the quality of air decreases, it shows that there are occupants in the building and ventilation system running at these points for maintaining good air quality in the building. Similarly the quality of air remains good while there are no occupants in the building and it has been observed that during air quality remains good during the night time and on weekends, thus ventilation system does not need to run in this period. It is noteworthy to observe that there is a sudden change (one point) in values at around mid day on 18/12/15 in the Figure 6.4. It has been observed that this type of behavior occurs when the window is opened by the occupants in the building and that's why from the air quality perspective the k-means have detected it as no occupant in building for around 15 minutes.

One other factor important factor is that the energy of the ventilation system can be saved in case if there are no occupants in the building. The red dotted rectangle is showing one case i.e. in the morning time at 19/12/15, the VAV signals show running of the ventilation system, but the other variables are indicating that there is no one in the building, as can be observed in the Figure 6.4. The reason behind this is the controller has hard coded time set for specific time slots. As in the current scenario, normally there are people in the building on Saturday's till

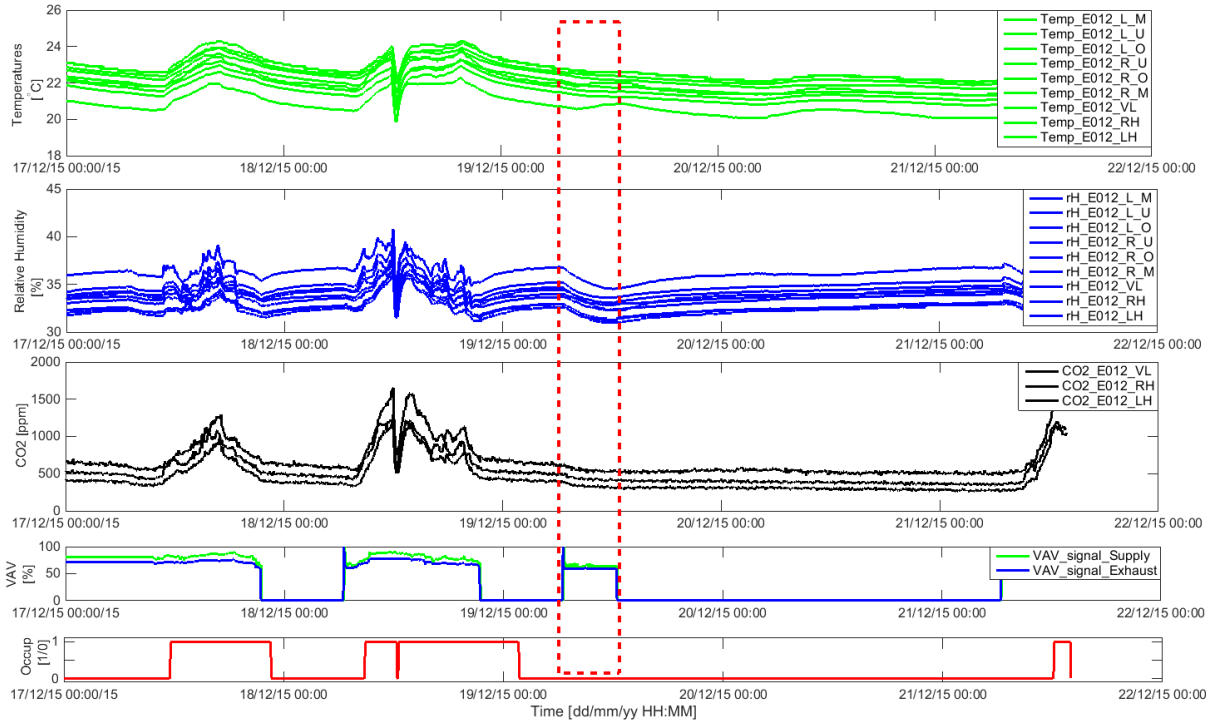


Figure 6.4: Possible areas for energy efficiency

12:00 o'clock mid day, and that's why the ventilation system is hard coded to 12:00 o'clock mid day. But on 19/12/15, there were no occupants as one the reason may be that there were no events in the room due to holidays (The data is taken from a seminar room where lectures are delivered). Therefore, energy can be saved if the ventilation system is controlled based on the occupants in the building.

6.1.2 Analysis of Chiller Operational Data

This section proposes a method for fault detection in operational data of energy systems. In order to validate the proposed model, data from absorption chiller is used. The method uses clustering algorithm for finding various types of behavior pattern of the energy system (chiller).

6.1.2.1 Methodology

The bag of words representation (BoWR) had been used to denote the different On cycles in the data. For finding the different behavior, hierarchical clustering had been used. The process of bag of word representation (BoWR) has already been discussed in Section 5.1.2 of the thesis. The hierarchical clustering partitions the On cycle data in such manner that the cycles representing same kind of behavior are gathered in one cluster, while the cycles representing different behaviors are put in different clusters.

In order to find the different patterns in the Operational data (On cycles), different tests were performed in consultation with the experts in the field. There are additional features added for getting better results. The temperature difference between the return and supply temperature sensors of each of the cycle had been used as a feature that are given as,

$$\Delta\text{Temp\_LT} = |\text{T\_LTre} - \text{T\_LTsu}| \tag{6.1}$$

$$\Delta\text{Temp\_HT} = |\text{T\_HTre} - \text{T\_HTsu}| \tag{6.2}$$

$$\Delta\text{Temp\_MT} = |\text{T\_MTre} - \text{T\_MTsu}| \tag{6.3}$$

The following Table 6.1 shows the different features used for the hierarchical clustering.

Table 6.1: Selected features for hierarchical clustering

Features	Descrition
ΔTemp_LT	Temperature difference of low temperature cycle
ΔTemp_HT	Temperature difference of high temperature cycle
ΔTemp_MT	Temperature difference of medium temperature cycle
Q6a_m3h	Flow in high temperature cycle
Q7_m3h	Flow in low temperature cycle
Q12_m3h	Flow in medium temperature cycle
Q6a_KW	Energy reading in high temperature cycle
Q7_KW	Energy reading in low temperature cycle
Q12_KW	Energy reading in medium temperature cycle

6.1.2.2 Bag of Words Representation (BoWR)

In order to represent the complete behavior of a cycle with different parameters taken under consideration, each sensor data is converted to the bag of word representation of 60 characters and put together in a 540 character representation, as can be seen in Figure 6.5.

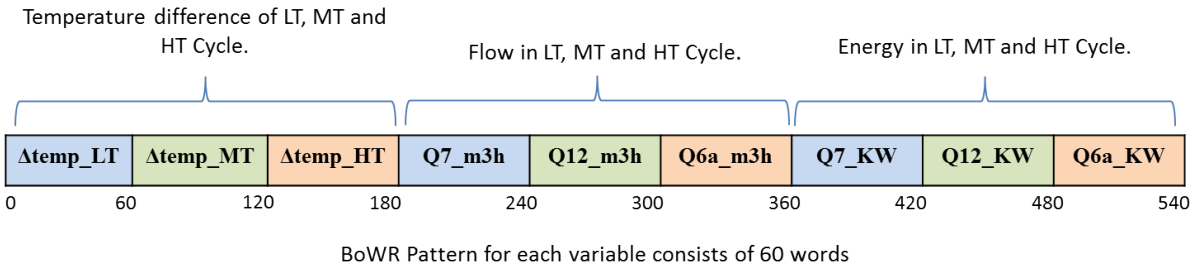


Figure 6.5: 540 character long representation of Cycle

As discussed in Section 5.1.2, the same process of conversion to BoWR will be used. All the required parameters will be converted to Z-Score before translation to symbolic aggregate approximation (SAX) symbols of the time series data. The value of M is taken as 60 for this research. The bag of words representation ($\mathrm{BoWR_i}$) for the i^{th} On cycle containing all features (given in Table 6.1) pattern can be defined as,

$$\mathrm{BoWR_i} = \{\mathrm{BoWR\_Sensor^i_1}, \mathrm{BoWR\_Sensor^i_2}, ..., \mathrm{BoWR\_Sensor^i_N}\}, \tag{6.4}$$

The $\mathrm{BoWR\_Sensor^i_N}$ of sensor N consists of a vocabulary set $\{w_1, w_2, w_3, ..., w_M\}$. The associated histogram vector $\mathrm{BoWR\_Sensor^i_N}$ for i^{th} cycle will be as following,

$$\mathrm{BoWR\_Sensor^i_N} = \left(\mathrm{V^i_1\ V^i_2\ V^i_3\ .....\ V^i_M}\right), \tag{6.5}$$

where N is representing the features selected for finding the different patterns in the chiller data given in Table 6.1. The V^i_j is the number of occurrences of w_j in the i^{th} cycle, i.e.

$$\mathrm{V^i_j} = \mathrm{Count_i(w_j)}, \tag{6.6}$$

here the subscript i in $Count_i$ refers to the i^{th} cycle.

In order to handle the different time length cycles a better idea is to normalize, i.e. use relative frequencies. With this in view, Equation 6.6 has been modified as Equation 6.7 given below, where N_i is representing the number of time ticks in the i^{th} cycle

$$\mathrm{V^i_j} = \frac{\mathrm{Count_i(w_j)}}{\mathrm{N_i}}. \tag{6.7}$$

6.1.2.3 Hierarchical Clustering

The hierarchical clustering technique groups the data over a different scales by creating a cluster tree called dendrogram. The dendrogram shows a multilevel hierarchy of clusters, where the clusters (groups) at one level are joined together as clusters at the next level. This property of

hierarchical clustering allows to decide the level of clustering that is the most appropriate for the task it is used for.

The bag of word representation (BoWR) for each cycle is clustered using the hierarchical clustering technique to find the different clusters of patterns in the data. Figure 6.6 shows the dendogram of the $BoWR_i$ given as input to the hierarchical clustering.

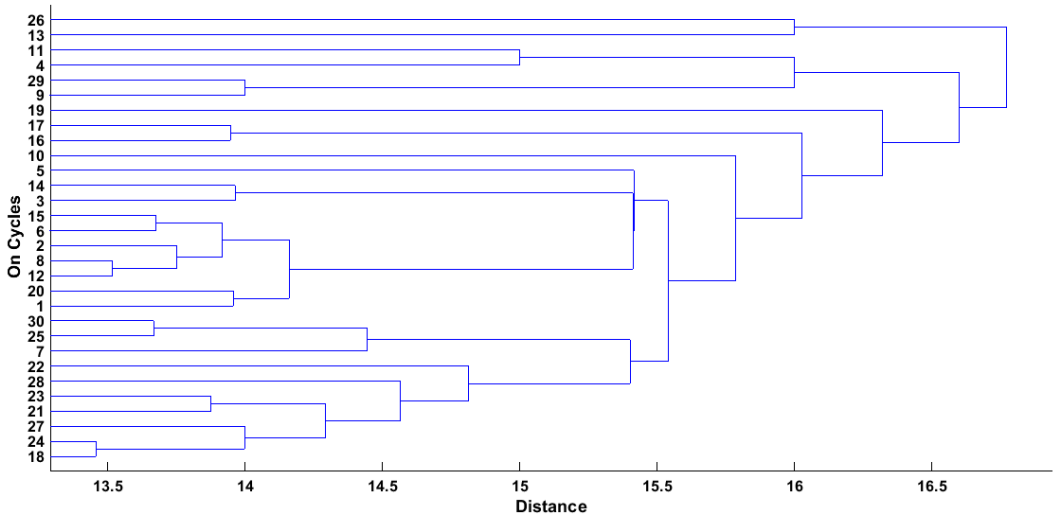


Figure 6.6: Dendrogram of bag of words representation (BoWR)

There are different techniques available to decide the best level or number of clusters for hierarchical clustering. One such method is the gap method [168]. Figure 6.7 show the gap distance against the number of clusters for the given $BoWR_i$ of On cycles and it can be seen that the data has best gap distance with 5 clusters. The result of the clustering algorithm gives better results when the difference between the intra-cluster elements of the cluster is as less as possible while the inter difference between the different clusters is as high as possible.

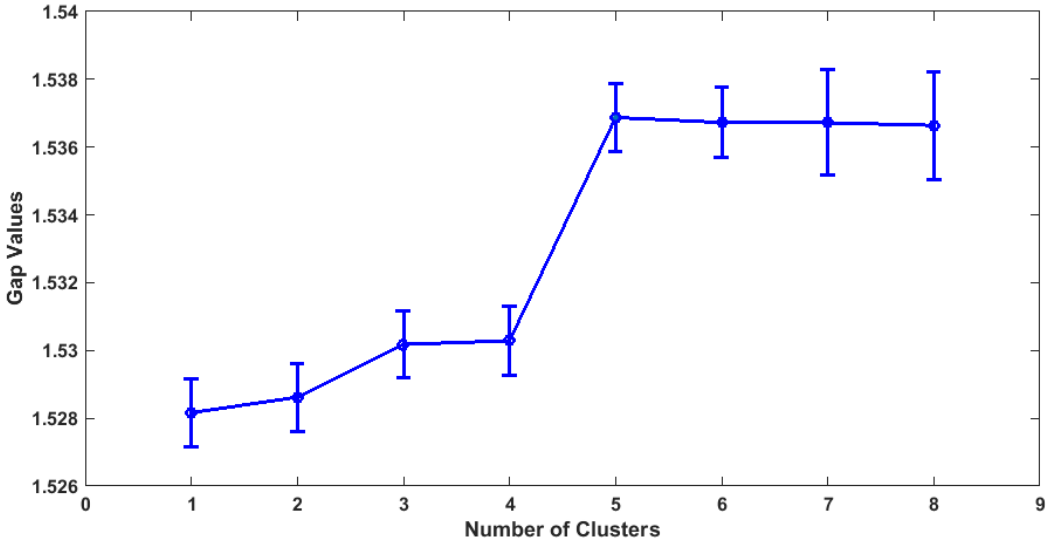


Figure 6.7: Gap distance of the BoWR data

Figure 6.8 shows the steps involved in finding the different patterns in the data of a water based chiller. It can be summarized as follows,

- Find the On/Off cycles of chiller using k-Means clustering algorithm as discussed in Section 4.1.2.2 of the thesis.
- Create a bag of word representation (BoWR)for all the On cycles. The BoWR will contain the selected features patterns.
- Use hierarchical clustering to group the similar cycles having same patterns.

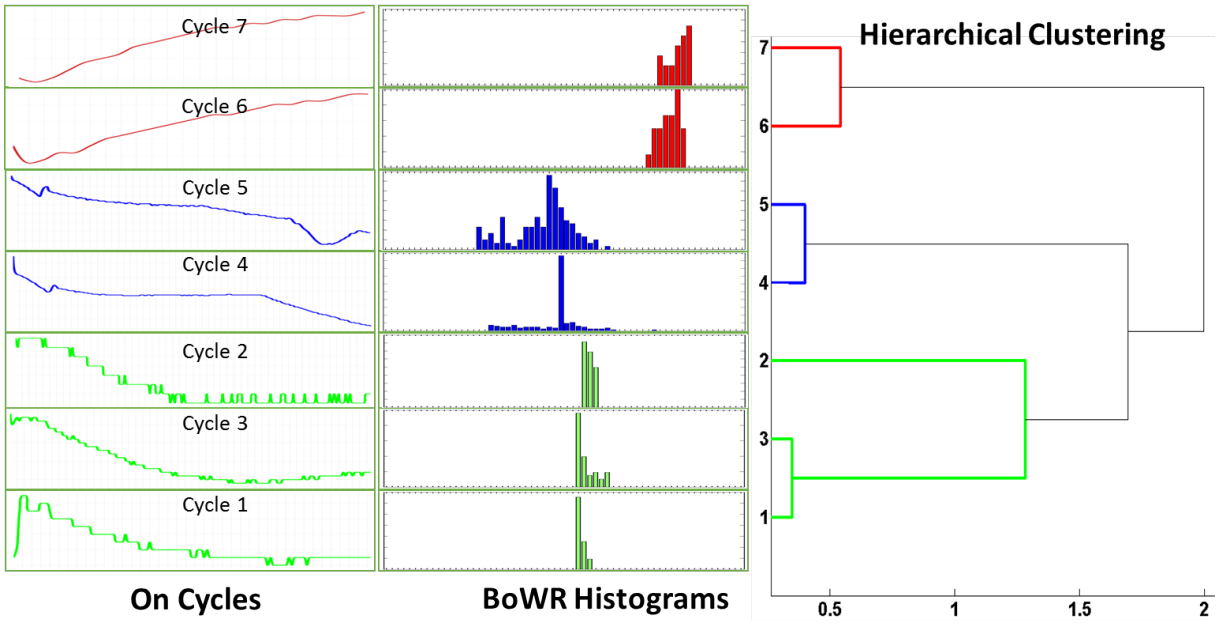


Figure 6.8: Methodology for finding fault patterns in chiller data

6.2 Diagnosis of Faults

This section discussed the methodology adopted for diagnosis of faults specifically in chillers. There are various rules proposed for diagnosing the faults in solar water based (absorption) chillers.The control unit implemented for the systems taken under consideration does not let the chillers to operate in such a way that can cause damage to them. Therefore, state of the art faults such as blocked duct in low temperature cycle are not found in the data as the control unit invokes the non-operational state of chillers in such cases, in order to prevent any damage.

It has been observed that during the monitoring of the systems, the control unit did not triggered any warning of the faulty operation. Therefore, in order to diagnose faults, those systems On cycles where the required cooling was no more possible because of the missing cooling capacity of the chiller are explored. In this research the cluster of On cycles where the average coefficient of performance (CoP) was low and running for longer time are further investigated for possible faults. To diagnose the reason for bad CoP, after a comprehensive discussion with experts in the field, the following additional parameters of operation had been integrated into the analysis:

- Evaporation pressure set point of absorbent($P_{Eva-setpoint}$).
- Low pressure of the system(P_{system}).
- Injection value status.

6.2.1 Types of Faults for Diagnosis

Figure 6.9 describes the internal structure of the chiller. There are two pressure zones in the system to work, as can be seen in the Figure 6.9, separated by a red dashed line. The valves

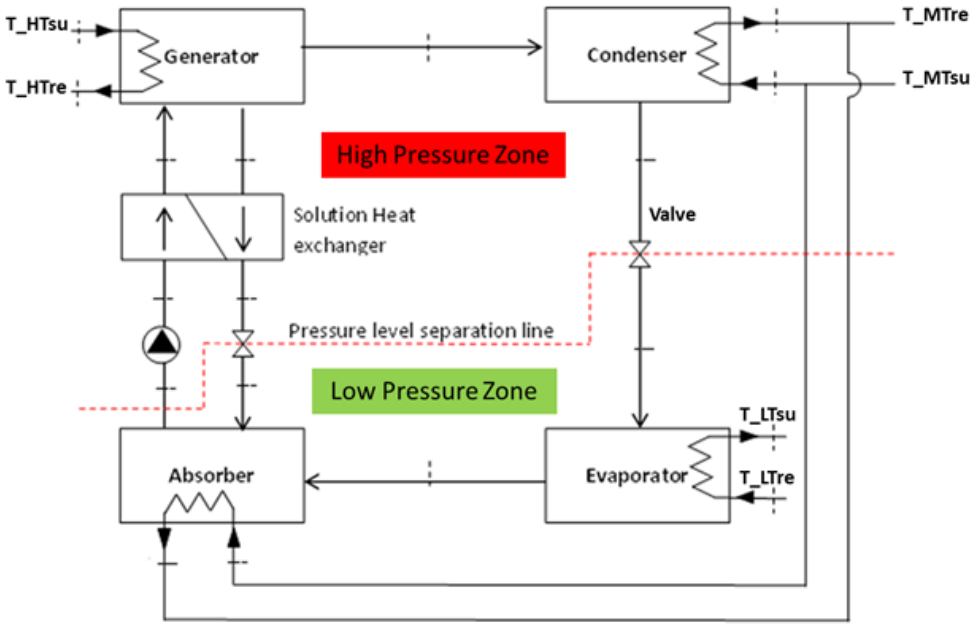


Figure 6.9: Architecture of the chiller

are the important meeting point for these areas as for maintaining pressure in both zones. It is required to have low pressure in evaporator as to help evaporate the absorbent (ammonia) that creates the cooling effect. For every absorbent there is pressure set point for evaporation ($P_{Eva-setpoint}$). It was eventually decided that the difference between the pressure set point for evaporation ($P_{Eva-setpoint}$) and the real value of the pressure in the low pressure part (P_{system}) of the chiller can be used to diagnose faults if exists in the system.

For diagnosing the reason behind the bad operation of the machine, the cluster of cycles having bad average CoP and longer operational period were further investigated. After different observations and suggestion by the experts the following conditions are selected for finding and diagnosing the different types of faults of chiller in the data, also shown in Figure 6.10

1. **LT Pump fault**: If the cooling load requirement is not fulfilled, then check the flow in the low temperature cycle. If there is no flow in the LT cycle ($Q6\_m3h = 0$), then this is hint for LT pump not in a working condition.

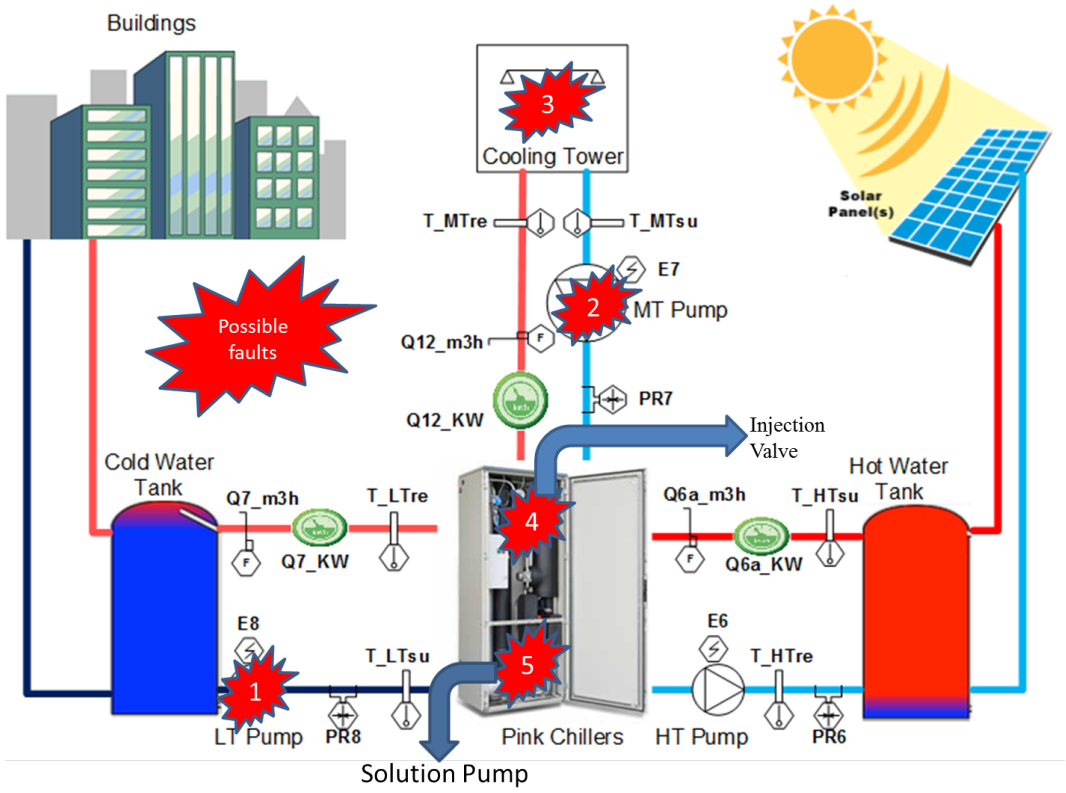


Figure 6.10: Faults in the System

2. **MT Pump fault**: It is similar with the previous point. In case of poor performance, if there is no flow in the MT temperature cycle ($Q12\_m3h = 0$) shows that the pump is not properly functioning in the medium temperature cycle.

3. **Cooling Tower Fault**: For this kind of faults, the difference of evaporation set point pressure of absorbent and low pressure of the system is required which can be calculated as,

$$\Delta P = P_{Eva-Setpoint} - P_{System} \tag{6.8}$$

   If $-1 < \Delta P < 1$ and $\Delta P \neq 0$ is the condition that defines the problem with cooling tower.

4. **Injection Valve fault**: If $\Delta P < -1$ denote the fault in injection valve.

5. **Solution Pump fault**: In case when $\Delta P > 1$ represent that the solution pump is in faulty state.

6.2.2 Algorithm for Fault Diagnosis

Algorithm 6.1 outlines our fault diagnosis strategy. The input is constituted by the clusters of cycles along with the CoP information and average operational period of cycles. The algorithm returns a variable $faulty\_cycle$ that may assume any of the values represented in Table 6.2 according to the five types of faults discussed in the previous section; absence of any fault is denoted by a zero.

Constant	value	Description
NO_FAULT	0	No Fault
LT_FAULT	1	Low Temperature pump Fault
MT_FAULT	2	Medium Temperature pump Fault
CT_FAULT	3	Cooling Tower Fault
IV_FAULT	4	Injection Valve Fault
SP_FAULT	5	Solution Pump Fault

Table 6.2: Fault Constants

Algorithm 6.1: Sensor fault diagnosis Pseudo Code

```
Input: On Cycles with Cluster information, Average_CoP_Cluster,
Average_Time_Cluster;
begin
Initialize: faulty_cycle_i := NO_FAULT, ∀i
for i = 1 : num_of_clusters do
    if Average_CoP_Cluster < 0.25 AND Average_Time_Cluster > 0.75(45 minutes) then
        for j = 1 : num_of_cycles_in_cluster_i do
            if Average(Q7_m3h_Cycle_j) = 0
                faulty_cycle_j := LT_FAULT
            else if Average(Q12_m3h_Cycle_j) = 0
                faulty_cycle_j := MT_FAULT
            else if Valve_Injection = On
                ΔP = P_Eva−Setpoint − P_System
                if ΔP > −1 and ΔP < 1 and ΔP ≠ 0
                    faulty_cycle_j := CT_FAULT
                else if ΔP <= −1
                    faulty_cycle_j := IV_FAULT
                else if ΔP >= 1
                    faulty_cycle_j := SP_FAULT
                end if
            end if
        end
    end
end
return faulty_cycle
end
```

6.3 Experiments and Results

The experiments had been performed on the data from water based chillers. The selected features were converted to BoWR. The On (operational) cycles are more appropriate for finding faults in the chiller, therefore, only On cycles are considered for clustering. The hierarchical clustering was applied with dynamic time warping (DTW) and proposed method of bag of words representation (BoWR) method. The comparison of the hierarchical clustering performance of the two methods was carried out with the help of cophenetic coefficients [169]. The cophenetic correlation is

technique that demonstrates the cluster tree strong correlation with the distances between objects in the distance vector. Table 6.3 shows the cophenetic coefficients with different hierarchical clustering methods for BoWR and DTW techniques. The BoWR has strong correlation with distance with all other objects in all clustering method. The best results for BoWR are attained with *Average* method for hierarchical clustering.

Table 6.3: Cophenetic Coefficients of Dynamic Time Warping (DTW) and BoWR

No.	Clustering Methods	BoWR	Dynamic Time Warping (DTW)
1	Average	**0.9897**	0.0375
2	Centroid	0.9851	0.037
3	Complete	0.9753	0.035
4	Median	0.9803	0.0363
5	Single	0.9848	0.0414
6	Ward	0.9835	0.0363
7	Weighted	0.9888	0.0368

The hierarchical clustering makes a clustering tree (dendrogram) that gives the option to select the level (cutoff) for clustering. The gap statistics [168] had been used to find the optimum number of clusters of cycles, depending on the gap between different clusters. As discussed in Section 6.1.2.3, the gap statistic gives the best gap distance with five clusters.

The cluster information of the five clusters are given in Table 6.4. The interesting patterns group is $Cluster_1$ and $Cluster_2$ as the average operation time in these clusters is greater than around 1 hour. For finding faults $Cluster_1$ patterns are more suitable as average coefficient of performance (CoP) of cycles in this cluster is 0.16 and average operational time of cycles is around 68 hours, thus showing the chiller performance is bad. The majority of the On cycles lie in the $Cluster_1$ as 98.75% of the cycles are in this cluster. $Cluster_1$ shows the cycles with normal operational behavior as the average coefficient of performance (CoP)is 0.54, while its average operational time of the cycles in cluster is around one hour. $Cluster_3$,$Cluster_4$ and $Cluster_5$ are representing the cycles with shorter operational time, as the machine is in transient phase, thus the patterns in these are different from normal operational behavior of the chillers and are not plausible.

Table 6.4: Cluster information of the five clusters with hierarchical clustering

Cluster_No	Percentage of Cycles in cluster (%)	Average CoP of On Cycles in cluster	Average Time of On cycles in cluster (hours)
$Cluster_1$	0.73	0.16	67.65
$Cluster_2$	98.75	0.54	0.95
$Cluster_3$	0.06	0.62	0.09
$Cluster_4$	0.34	0.87	0.07
$Cluster_5$	0.12	0.7	0.06

In order to see the details of how the hierarchical clustering has grouped the elements together, the Figure 6.11 show the 3D graph of the $\Delta Temp$, energies, flows of the normal behavior and faulty behavior of the chiller. The data is normalized with min-max normalization before visualization. The 3D graph displays the elements of the two clusters of the system that are grouped together

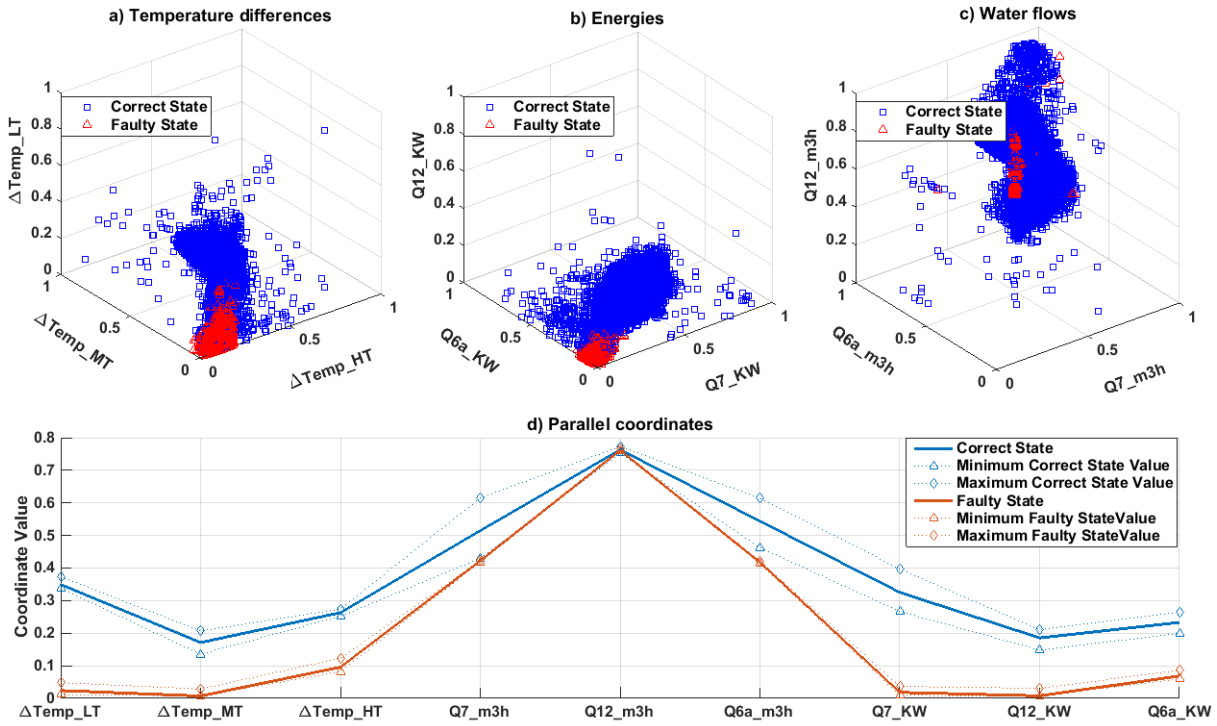


Figure 6.11: (a) 3D graph of $\Delta Temp$ of the HT, MT and LT cycle. (b) 3D graph of energies in the HT, LT and MT cycle. (c) 3D graph of flows in the HT, LT and MT cycle. (d) Parallel coordinates of the hierarchical clustering

with hierarchical clustering. The groups can be seen with different color of elements. The red color represent the data in the faulty state while the blue color represents the data in normal behavior cluster.The parallel coordinates also draws a good view of the grouping. The parallel coordinates show the mean, maximum and minimum of normal and faulty patterns of the data. It is noteworthy to mention that the cluster elements are not making any cluster that can be seen, as the reason is the there is flow recorded in the faulty pattern as can be observed in the Figure 6.11(c), but the difference between the faulty and normal patterns can be seen in the energies and temperature difference figures. The same has been shown in the parallel coordinate graph as can be seen in Figure 6.11(d).

For further investigation of the cycles behavior in $Cluster_1$, the Figure 6.12 has been drawn to show the behavior pattern of one of the On cycles in $cluster_1$. The 3 graphs in Figure 6.12 display the temperature difference ($\Delta Temp\_LT, \Delta Temp\_HT, \Delta Temp\_MT$), flows ($Q7\_m3h, Q6a\_m3h$, $Q12\_m3h$) and energy meter reading ($Q7\_KW, Q6a\_KW, Q12\_KW$) in low temperature cycle, high temperature cycle and medium temperature cycle of the chiller operational cycle. The x-axis display the time in minutes for the On cycle. In each graph the values are represented with the intensity of the color given in the form of a vertical color code bar at the right hand side of each plot. It can be observed form the Figure 6.12 that at 30 minutes the $\Delta Temp\_MT$ becomes zero showing that the cooling tower is not operating normal, causing no change int the temperature of medium temperature cycle. It is also important to note that the flow variable ($Q12\_m3h$) for MT cycle is showing flow throughout the cycle. Due to this effect the $\Delta Temp\_LT$ has also started decreasing and at around 80 minute the cooling has been stopped by the chiller. At the same time the deriving heat ($\Delta Temp\_HT$) has been provided to the chiller but the chiller

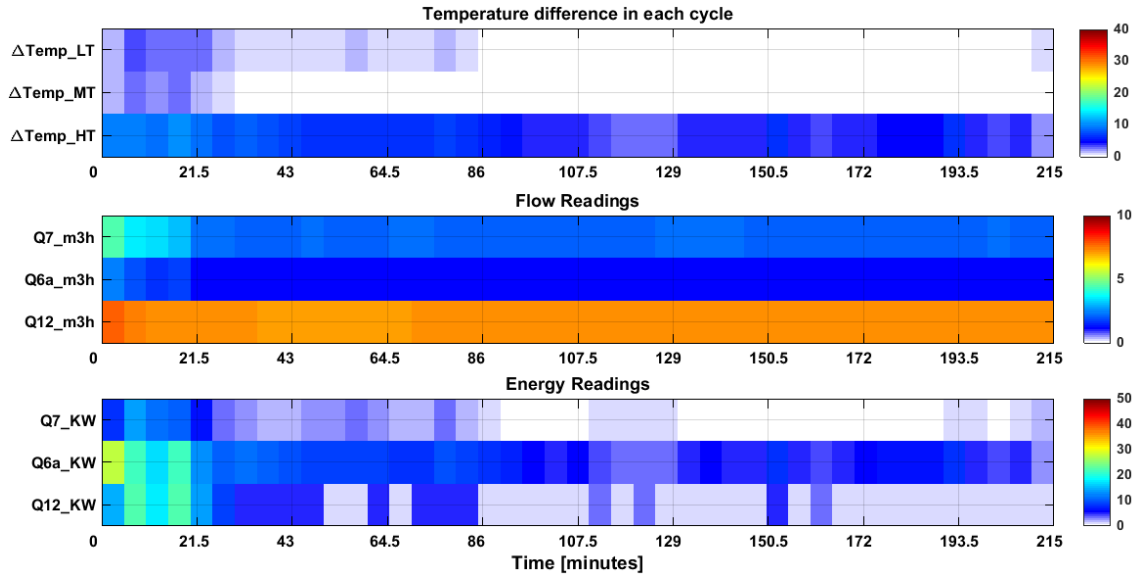


Figure 6.12: On cycle pattern showing low performance (faulty) operation of chiller

is not able to match the cooling load, thus showing lower coefficient of performance. This pattern of the chiller is giving a clue about the faults in the chiller that need to be diagnosed.

Furthermore, to support the argument that $Cycle_1$ is representing group of On cycles having bad performance, the histogram of coefficient of performance of On cycles grouped in $Cluster_1$ are shown in Figure 6.13. The histogram shows that 80% of the On cycles have CoP less than 0.2, representing low performance of the chiller.

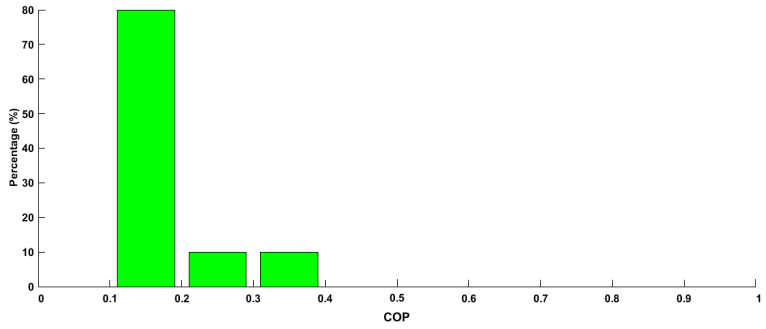


Figure 6.13: Histogram of On cycle's associated with cluster of faulty operation of chiller

The same procedure has been used to see the behavior pattern of one of the On cycles in $cluster_2$. The a-axis display the time in minutes for the On cycle. In each graph the values are represented with the intensity of the color given in the attached color bar.

It can be observed form the Figure 6.14 that the duration of the On cycle is 195 minutes. The temperature difference parameters show that there is cooling in the LT cycle as $\Delta Temp\_LT$ is representing it with time. The effect can be seen in MT cycle as well. At the same time, HT cycle shows that the constant driving heat was provided to the chiller. It is also important to mention that the flow variable display the flow in all the three cycle, whereas, same has been observed for the energy parameters. This behavior pattern shows the normal operation of the chiller.

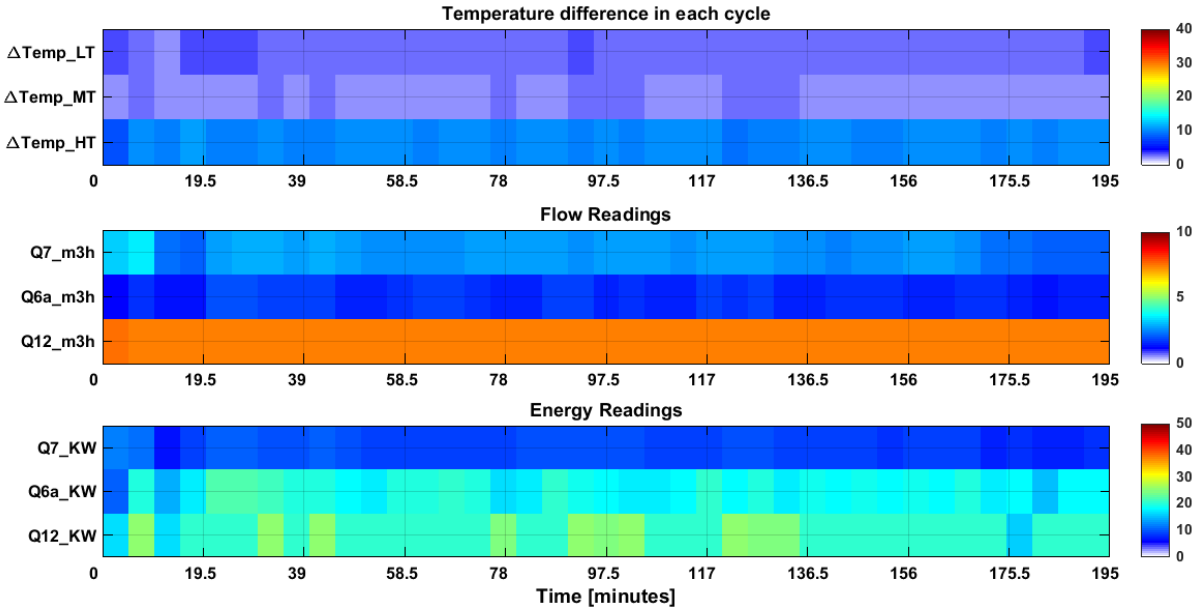


Figure 6.14: On cycle pattern showing normal operation of chiller

The histogram of coefficient of performance (CoP) of On cycles grouped in $Cluster_2$ is shown in Figure 6.15 in support of the argument that $Cycle_2$ is representing the group of On cycles corresponding to the normal performance of the chiller. The histogram shows that the chiller is performing with CoP between 0.4 to 0.8. Thus, representing the normal behavior of the chiller.

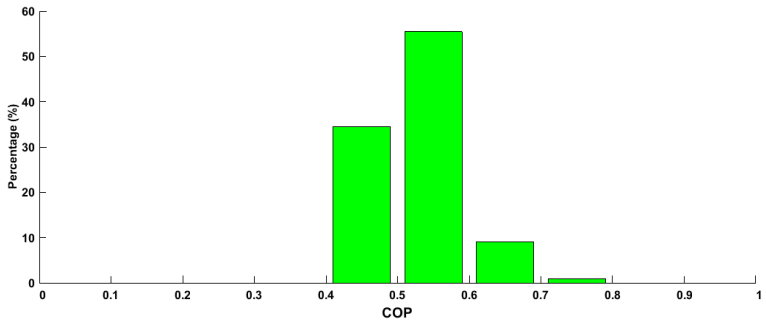


Figure 6.15: Histogram of On cycle's associated with cluster of normal operation of chiller

For diagnosis of faults the algorithm suggested had been tested with the data from all the clusters. The elements of $Cluster_1$ were demonstrating low performance, therefore the cycles in this cluster are further investigated for faults with the additional parameters taken during diagnosis.

The detailed analysis of the water based chiller was carried out with the help of algorithm suggested in Section 6.2.2 for the five faults discussed in Section 6.2.1. This can be readily observed from Figure 6.16. The x-axis in the figure shows the time of operation for the chiller from *May* 2014 to *May* 2015. The line represented with working period for the chiller is showing where the chiller was operational in the given time. It can be observed from the Figure 6.16 that there are solution pump and cooling tower faults diagnosed in January 2015 represented with red diamonds symbols. After verification from the technical manager of the site, it was found out

that the chiller was not performing well due to the damage to cooling tower fan. These faults detected by the proposed method were real which the monitoring system failed to detect.

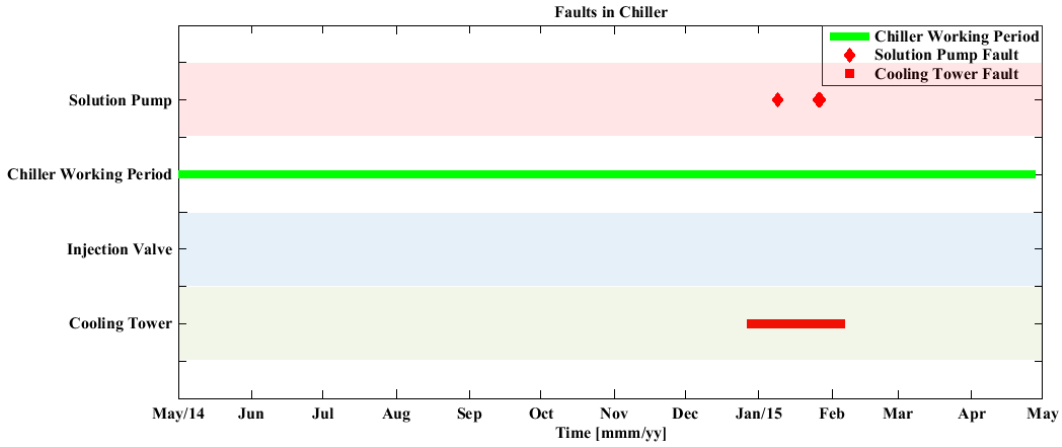


Figure 6.16: Faults in the operation of the chiller

7 Conclusions and Outlook

This chapter summarizes the contributions of the work done in this thesis and suggests future directions. The energy demands of the world had been increased tremendously in last couple of decades and predicted to escalate by almost 50% from the year 2009 till 2035 [1]. The share of the buildings is significant in the total energy consumption, which is 21% of total energy requirements at the world level [1]. The building energy consumption has major shares in the developed countries and regions, such as USA and EU, where the energy demands in buildings are around 41% and 40% of the total energy consumption, respectively [2, 3]. Moreover, even in the underdeveloped countries like Pakistan, buildings have significant figures in the electricity consumption which is about 44% [170]. Furthermore, the energy systems in buildings such as heating, ventilation and air conditioning (HVAC) share the key contribution in buildings electricity consumption as HVAC systems are required to maintain the desired comfort conditions inside the buildings. Therefore, automated fault detection and diagnosis in energy systems of the buildings provides a great potential for energy savings in the buildings sector.

There are various different methods that can be helpful in detection and diagnosis of faults, thus helps in reducing energy consumption. Data mining is one of the prominent research area that can help in automatically detecting the different patterns in buildings data; an example is clustering [12, 13, 14]. The process of automatically finding various patterns in the data can make the analysis easy, feasible and less labor extensive [12, 13, 15]. After the detection of faults, in the operational behaviors of the energy systems, it is required to diagnose the detected faults. The diagnosis of faults will help to know the reason behind the detected fault. Subsequently, actions can be taken to handle these issues, thus will help in saving cost and reduction in energy consumption.

The main objective of the research was to develop a process model and framework for the automatic detection and diagnosis of the various types of faults in the energy system of the buildings, emphasizing on less dependence on experts knowledge during the analysis process. The proposed work flow will be beneficial for those users who need to automatically evaluate the energy systems of the buildings and find various types of faults in the stored data without involving experts e.g. building facility management service providers. Additionally, the focus of this research was to use data analytics and machine learning methods for developing a work flow for automatic detection and diagnosis of various anomalies in the recorded data, and faults in the operation of energy systems in buildings, with minimum as possible support from the experts in the field.

7.1 Contributions

The main contribution of this dissertation was to develop work flow using a robust set of algorithms for automated data quality check and data analysis for fault detection and diagnosis in the energy system of buildings, with little configuration efforts. The proposed work flow method is more appropriate for the analysis of the energy systems in general. The first phase of the work flow was the data sanitation, where solution for various issues had been given, such as automatic duty cycle detection, validation of data using rule of first principles and visualizations, limit checks, outlier detection, and missing data imputation. In the second phase the detection of different anomalies or noise added to the data, during recording of the data had been discussed. In the last phase the various types of faults in the operations of energy systems had been targeted and solution for automatic detection and diagnosis have been proposed for energy systems in buildings.

Table 7.1 show the different methods used in the work flow for detection and diagnosis of faults in energy system of buildings were selected taking in view the effort it takes to analyze large amount of data with less availability of additional information, that can also be called as meta data. Furthermore, the requirement of domain knowledge is required to be less as well. Table 7.1 describes the methods that required little or no domain knowledge.

Table 7.1: overview of the data analysis methods used in the work flow for detecting and diagnosis faults automatically

Method/View	Configuration Required	Domain Knowledge Required
Method: checking data,access and availability	No	No
View: calendar view for data availability	No	No
Method: k-means clustering for duty cycle detection	No	No
Method: z-score and expectation maximization for outlier detection	No	No
Method: interpolation for gap sanitation	No	No
Method: setting process limits	Yes	Yes*
Method: regression for sanitation of large gaps	No	Yes
View: histograms for on-state data only	No	Yes*
Method: first principle (energy balance, CoP)	Yes	Yes
Method: Anomaly detection	No	No
Method: Fault detection	No	Yes*
Method: Fault diagnosis	No	Yes
*Requires basic information typically found on data sheet		

The operational data of various energy systems have been used for the purpose of the validation of the proposed methodology such as solar based adsorption chillers, solar based absorption chillers, and air handling unit. The data was recorded and stored in JEVis. Furthermore, for analyses of the recorded data of the energy systems performance, a unified monitoring procedure developed in the framework of IEA SHC Task 38 [16] had been used. For experimentation and validation of the purposed methodology, the data from various solar based energy systems have been used. The details of the systems can be seen in Table 7.2, where system names along with the time period and number of sensor are given.

Table 7.2: Details of the energy systems used for experimentation

S.No	System Name	From	To	Number of sensors used for analysis
1	Absorption chiller (Machine 1)	June 2013	March 2014	l8
2	Absorption chiller (Machine 2)	June 2013	April 2015	18
3	Absorption chiller (Machine 3)	July 2013	April 2015	18
4	Absorption chiller (Machine 4)	January 2014	April 2015	18
5	Absorption chiller (Machine 5)	March 2013	April 2015	18
6	Adsorption chiller	January 2009	December 2012	26
7	Ventilation system	December 2015	February 2016	37

The duty cycle detection had been done using supervised learning methods such as k-nearest neighbors (KNN), support vector machines (SVM) and random forest trees. Different tests were performed and it was found out that random forest had shown the best results among the three methods as it was also not sensitive to the missing values. The KNN has shown that it was sensitive to noise, but has shown good results when values were filled with interpolation method. The SVM method performed better than KNN but performed below random forest trees.

Furthermore, in order to have a robust method for automatic detection of the duty(operational) cycle, the behavior of the different parameters such as temperatures, flows, energy readings and pressures of various energy system were analyzed. It has been observed that the data had two different patterns showing the operational (On) and non-operational (Off) state of the energy systems. Therefore, on the basis of this assumption that the energy system behavior pattern were different in the two states, a method had been proposed using k-means clustering algorithm that clustered the data in two clusters representing Operational (On) and non-operational (Off) cycle. The Figure 7.1 shows the accuracy results of k-means algorithm applied on various energy systems. The results of the k-means were quite good, except for absorption chiller_3, where accuracy was around 95%, but, still in acceptable range. One of the main reason for small errors in accuracy was due to the fact that k-means was not accurately detecting state information for the transient phase of the energy system. Furthermore, k-means was also used to find the occupants in the building using the ventilation system data.

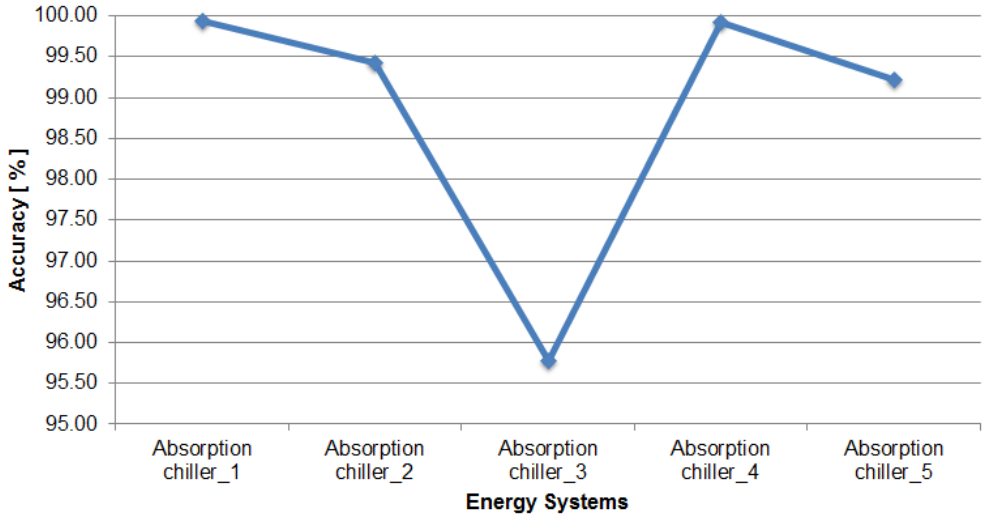


Figure 7.1: k-means accuracy graph

Moreover, the problem of gaps created in the data were handled using interpolation and regression methods, depending on the length of the gap. After various different experiments on data it had been observed that for short gaps of missing, interpolation method has produced better results as compared to regression. While, the results from experiments had shown that for longer gaps, regression has better results. Moreover, validation of data was suggested before detailed analysis as it saved time rather than analyzing data that did not represent the behavior of the desired energy system. Therefore, first principles and visualizations were proposed to validate the data with first hand knowledge. For limit check, histogram visualization was suggested. Figure 7.2 shows the ratio of different types of sensors that were found having values not in the acceptable range, using histogram method. The histogram visualizations showed the the pressure sensor had been observed violating the acceptable limit checks with 17%, followed by temperature sensors 12%.

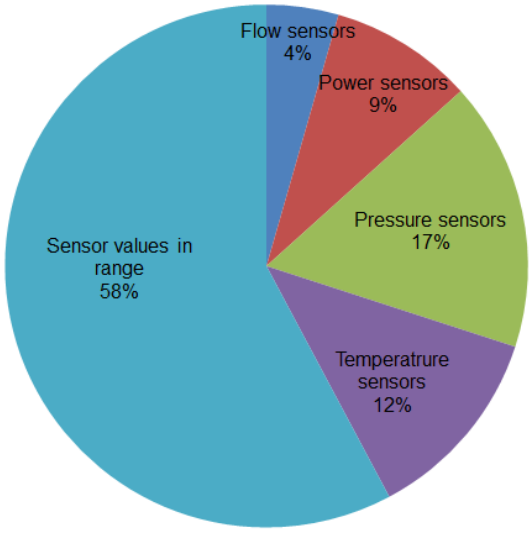


Figure 7.2: Limit check violation by sensors detected with histogram

As the behavior pattern of the energy system had been varying in two states i.e.operational (On) and non-operational state (Off), therefore this information had been used for automatic detection of outliers. The outliers can be defined as those data point that do not correspond to the expected normal behavior of the energy system. So a method for automatic outlier's detection had been proposed using z-score normalization based on the duty cycle (On/Off) state of the energy system, and subsequently, clustering the data by expectation maximization clustering algorithm. The duty cycle based z-score normalization helped in highlighting the outlier's for clustering algorithm to cluster these points away from the normal behavior of the data. Thus, helped in finding the invalid data points in the data. The results of proposed method were compared with other state of the art methods and it was found out that the proposed cycle based z-score method was able to detect outliers more accurately.

There were various types of anomalies in data of sensors recording, as sensors usually are deployed in fields for monitoring. A method had been proposed to automatically detect anomalies patterns in sensor data, that is recorded in energy systems of the building. A crucial issue with monitoring data in energy systems was the data quality of the recorded data, due to problems in commissioning, transmission, storage or the sensor itself, the data was often not feasible for further analysis, thus impairing the ability to detect faults in the operation of the energy sys-

tems. The focus lies on finding a method that covers all different possibilities of error patterns and has a high degree of automation so that it requires close-to-none configuration. The data is subjected to k-means clustering for the unsupervised classification of On/Off cycles and then transformed to a symbolic representation by applying the Symbolic Aggregate Approximation (SAX) method. The SAX symbols are converted to a Bag of Words Representation (BoWR) for each On/Off cycle. A pairwise comparison of the latter is carried out for the automated detection of anomalies patterns. The proposed method showed interesting results par rapport the reference methods. We used three types of reference methods. The comparison of the reference methods (histogram, ARIMA and artificial neural network (ANN)) with propose bag of word representation (BoWR) method. The results for *Constant* pattern can be seen in Figure 7.3. It has been observed that histogram method has the best *Probability of detection* with 99.83% for *Constant* faults. Moreover, artificial neural networks (ANN) has achieved the hight *True Alarm Rate* with 99.9%, while the highest *Accuracy* has been achieved by BoWR with 98.77% for *Constant* faults. (*True Alarm Rate* is equal to 1-*False Alarm Rate*.)

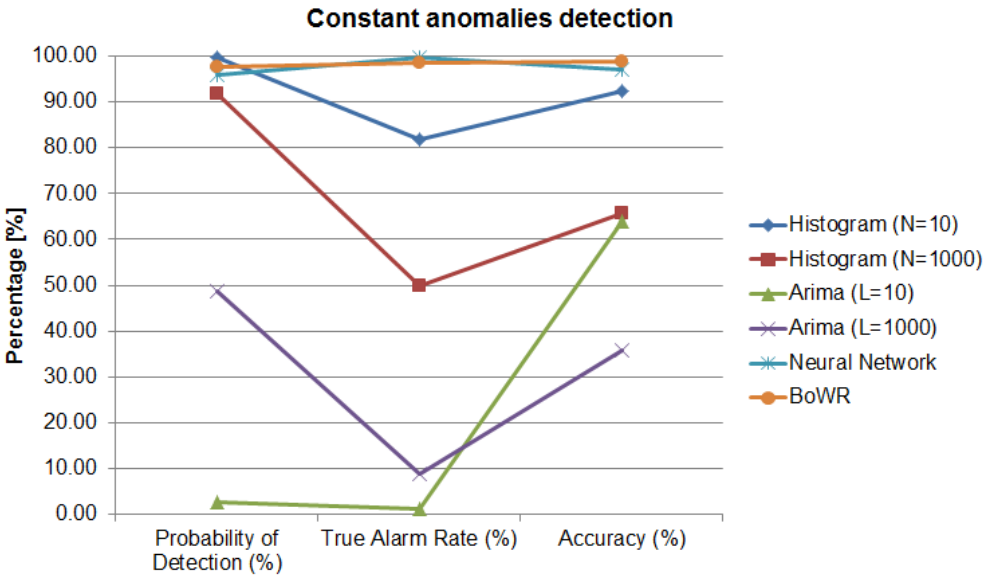


Figure 7.3: *Constant* anomalies detection comparison

The Figure 7.4 shows the results of the comparison for the *Noise* anomalies detection using the proposed BoWR method with the other state of the art method. The BoWR dominates in all the comparison parameters for detection of *Noise* faults. The BoWR had achieved the highest *Probabilty of detection* with 99.84%, maximum *True alarm rate* of 98.84% and best *accuracy* with 99.44% for detection for *Noise* faults as can be seen in Figure 7.4. Out of these rule-based (Histogram) methods can be highly accurate, but their accuracy was depending critically on the choice of parameters. Methods based on time series analysis (ARIMA) incurred more false positives than the other methods. The learning methods (ANN) can be cumbersome to train but can accurately detect and classify faults. In fact, in the absence of noise, the example ANN reference method performed at par with our BoWR method. Still the unsupervised nature of our proposed method and the requirement of only the On/Off information make it superior to others. Not to forget the fact that even the On/Off information is automatically detected by using the k-means clustering algorithm. All this notwithstanding, the BoWR method was heavily dependent on the correct detection of On/Off state by k-means.

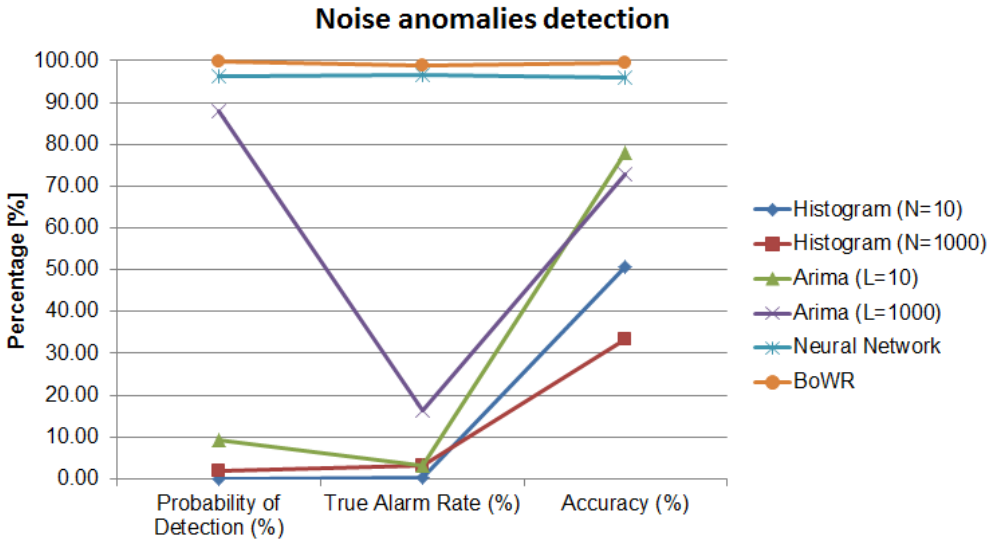


Figure 7.4: *Noise* anomalies detection comparison

The analysis of the energy systems was done, to be able to detect and diagnose faults automatically in energy systems of the building. During the analysis of the Ventilation system, it was found out that the ventilation system was running even when there were no occupants in the building. Therefore, energy consumption of the ventilation system can be reduced, if the ventilation system control is managed automatically by using the occupants in buildings information. Furthermore, in order to automatically detect faults in the operation of the energy systems and to find various patterns in the energy system operation, a bag of words representation (BoWR) with subsequent hierarchical clustering had been proposed. The method had used the duty cycle detection information by the k-means clustering algorithm. The operational (On) cycles were of greater importance for finding the performance of energy systems and further in detecting and diagnosing faults. Additional features were suggested for achieving better results of clustering algorithm. The hierarchical clustering used the BoWR of the On cycles of each feature for finding the various operational patterns of the energy system. The hierarchical clustering technique grouped the data over a different scales by creating a cluster tree which is also called dendrogram. The dendrogram showed a multilevel hierarchy of clusters, where the clusters (groups) at one level were joined together as clusters at the next level. This property of hierarchical clustering allowed to decide the level of clustering that is most appropriate for the task it was used. In order to decide the best level or optimum number of cluster, the gap statistics had been suggested. The cluster that had the duty cycles with less average performance and longer duration was considered to be possible area of faults.

After clustering the data, additional information was added to each cluster such as average coefficient of performance (CoP) of the cluster and average time of On duty cycles in a cluster. Furthermore, additional parameters of operation had been integrated into the analysis, enabling to diagnose these faults after comprehensive discussion with experts in the field. An algorithm using various rules had been suggested for the diagnosis of faults in the operation of the energy systems. The algorithm had targeted five different types of faults of water based (adsorption) cooling chiller systems i.e.,

- Low temperature (LT) pump not proper functioning.

- Medium temperature (MT) pump not proper functioning.
- Cooling tower not performing functionality.
- Injection valve fault
- Solution pump fault

The algorithm had been able to detect the three faults in the data that was under observation. The detected faults were cooling tower fault, solution pump fault and injection valve fault. The detected faults were cross checked with the logbook maintained at the site of each energy system for validation. It was confirmed that the problem were correctly diagnosed after comparison with diagnosed faults with the suggested rules.

7.2 Future Work

This section discusses the further development and improvement of various sections of the proposed work flow for automatic detection and diagnosis of faults in the energy systems of the buildings. The main point that need improvement can be summarized as following

- The advancement in the data acquisition is required that includes the current data integration and data fusion capabilities,
- The conversion of the raw data into useful information
- The increase in algorithm efficiency.
- Suggestion for automating the expert knowledge processes.

The International energy agency (IEA) has launched a technology initiative known as "IEA Solar Heating and Cooling Programme (SHC)". Within this implementing agreement IEA SHC Task 48 "Quality Assurance & Support Measures for Solar Cooling Systems" was one of the research topics. The IEA SHC Task 48 (Activity B6:"Report for self-detection on monitoring procedure") explains various faults found in different solar based energy systems [111]. They have categorized the faults in five different categorizes i.e. faults due to the bad design, sensor anomalies, control strategies problem, components defect, and components aging causing degradation in performance. It has been reported in the IEA SHC Task 48 that the detection of faults due to bad design are very difficult to diagnose. As the research work done in this thesis focused more on the detection of faults that are due to components, sensors anomalies detection, and faults in control strategies faults such as fault detected in the ventilation system, where the ventilation system was operational even with no occupants in the building. Therefore, additional research in the direction of detecting faults due to faulty design and performance degradation because of the aging factor can be added to current work flow.

One of the primarily issue that is noticed from the personal experience which needs improvement is the data acquisition phase, where the chances of the data corruption is quite high. The process of data acquisition should be automated having the capability of detecting anomalies at the run time in data with least configuration changes. The current simple monitoring systems are capable of validating data with simple techniques such as limit checks and first principles, and outlier's

detection methods but still more advance tools and techniques are required that do not need any specific information or configuration changes about the system. These tools should be able to automatically convert the raw data in to useful information for the analysis phase.

As for this thesis the data from the solar based systems (chillers and air handling unit) had been used for experimentation and testing the methodology. The adopted methodology need to be tested on the other conventional energy systems data. Secondly the IEA SHC Task 38 monitoring policy of International Energy Agency had been used for naming conventions and for finding the various performance evaluation parameters of different energy systems. The performance evaluation is helpful in detection of faults, but for diagnosis of faults a generic framework is required.

The suggested automatic duty cycle detection method using k-means in the proposed work flow had not been able to detect the duty cycle accurately during the transient phase of the energy system, as either the energy system is starting operation or about to end operation. Therefore, detection of transient phase is more complex task. Thus, a robust algorithm for automatically detection of duty cycles is required having accurate results specifically in the transient phase. The transient phase is also important for detection and diagnosis of several faults in the energy system of the buildings.

One of the common reasons in the existence of faults in the energy systems of buildings is the aging of the equipment. Currently this research work does not cover these kinds of faults. Therefore, more effort is required in this side, which can be added to the proposed work flow.

The fault diagnosis method proposed in this research is limited for one type of energy system i.e. water based (absorption) chillers. Additionally, the rules that were proposed were able to detect few faults that were observed during the analysis of the operational data for finding various patterns in the data. So, a robust method for diagnosis is required to be added in the proposed work flow that should be able to diagnosis faults automatically.

The suggested algorithms in the proposed work flow for energy systems of the buildings were not tested for algorithm efficiency in terms of processing and memory. The reason is that, for the existing scenario of the recorded data, the proposed methodology algorithms were able to produce results in acceptable time and memory. But in future and for various scenarios it is required to test the efficiency in terms of processing and memory for the used algorithms in the proposed work-flow.

References

[1] US Energy Information Administration (EIA), "International energy outlook 2011," Tech. Rep. DOE/EIA-0484, USA, 2011.

[2] Buildings Performance Institute Europe (BPIE), "Europe's buildings under the microscope. a country-by-country review of the energy performance of buildings," tech. rep., Brussels, 2011.

[3] U.S. Department of Energy, "2011 buildings energy data book," tech. rep., Building Technologies Program, Energy Efficiency and Renewable Energy., March 2012.

[4] Intentional Energy Agency (IEA), "Electricity information," tech. rep., Paris, 2009.

[5] M. Ellis and E. Mathews, "Needs and trends in building and hvac system design tools," *Building and Environment*, vol. 37, no. 5, pp. 461–470, 2002.

[6] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and buildings*, vol. 40, no. 3, pp. 394–398, 2008.

[7] M. S. Breuker and J. E. Braun, "Evaluating the performance of a fault detection and diagnostic system for vapor compression equipment," *HVAC&R Research*, vol. 4, no. 4, pp. 401–425, 1998.

[8] J.-L. B.-K. Mohammad Mourad, "A method for automatic validation of long time series of data in urban hydrology," *Water Science and Technology*, vol. 45, pp. 263–270, Mar. 2002.

[9] N. O. a. A. A. US Department of Commerce, "Handbook of Automated Data Quality Control Checks and Procedures," Tech. Rep. NDBC Technical Document 09-02, National Data Buoy Center, Mississippi 39529-6000, Aug. 2009.

[10] S. Sun, J.-L. Bertrand-krajewski, A. Lynggaard-Jensen, J. V. D. Broeke, F. Edthofer, M. D. C. Almeida, Á. S. Ribeiro, and J. Menaia, "Literature review for data validation methods," Tech. Rep. PREPARED 2011.019, PREPARED: Seventh Framework Programme, 2011.

[11] A. Mattarelli and S. Piva, "EN 15316 calculation methods for the generation sub-system: The influence of input data on the results," *Energy Procedia*, vol. 45, pp. 473 – 481, 2014. ATI 2013 - 68th Conference of the Italian Thermal Machines Engineering Association.

[12] C. Miller, Z. Nagy, and A. Schlueter, "Automated daily pattern filtering of measured building performance data," *Automation in Construction*, vol. 49, Part A, pp. 1–17, Jan. 2015.

[13] F. Iglesias and W. Kastner, "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns," *Energies*, vol. 6, pp. 579–597, Jan. 2013.

[14] B. Narayanaswamy, B. Balaji, R. Gupta, and Y. Agarwal, "Data Driven Investigation of Faults in HVAC Systems with Model, Cluster and Compare (MCC)," in *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, BuildSys '14, (New York, NY, USA), pp. 50–59, ACM, 2014.

[15] J. Lin and Y. Li, "Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation," in *Scientific and Statistical Database Management* (M. Winslett, ed.), no. 5566 in Lecture Notes in Computer Science, pp. 461–477, Springer Berlin Heidelberg, 2009.

[16] A. Napolitano, W. Sparber, A. Thür, P. Finocchiaro, and B. Nocke, "Monitoring Procedure for Solar Cooling Systems," Tech. Rep. IEA Task 38, International Energy Agency, Oct. 2011.

[17] X. Jin and J. Han, "Expectation Maximization Clustering," in *Encyclopedia of Machine Learning* (C. Sammut and G. I. Webb, eds.), pp. 382–383, Springer US, 2011.

[18] Invensys Building Systems, *HVAC Controls Introduction*. USA, 2001.

[19] P. Srikhirin, S. Aphornratana, and S. Chungpaibulpatana, "A review of absorption refrigeration technologies," *Renewable and Sustainable Energy Reviews*, vol. 5, pp. 343–372, Dec. 2001.

[20] T. Berntsson and P. Holmberg, "Alternative working fluid in heat transformers," *ASHRAE in of Transactions American Society of Heating Refrigerating and Air Conditioning Engineers*, vol. 96, no. 1, pp. 1582–1589, 1990.

[21] H. Perez-Blanco, "Absorption heat pump performance for different types of solutions," *International Journal of Refrigeration*, vol. 7, pp. 115–122, Mar. 1984.

[22] GBU mbH, "Adsorption chiller nak," tech. rep., 1999.

[23] H.-M. Henning, *Solar-assisted air conditioning in buildings: a handbook for planners*. Springer Verlag Wien, 2007.

[24] ASHRAE Handbook, *HVAC systems and equipment*. American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta, GA, 1996.

[25] P. Palensky, "The JEVis System – An advanced Database for Energy-related Services," ACTA Press.

[26] J. E. Braun, "Reducing energy costs and peak electrical demand through optimal control of building thermal mass," *ASHRAE Transactions*, pp. 264–273, 1990.

[27] M. McKellar, "Failure diagnosis for a household refrigerator.," Master's thesis, School of Mechanical Engineering, Purdue University, West Lafayette, Indiana., 1987.

[28] J. Hyvärinen and S. Kärki, "International energy agency building optimisation and fault diagnosis source book," tech. rep., Technical Research Centre of Finland, Laboratory of Heating and Ventilation,Espoo, Finland., 1996.

[29] A. Dexter and J. Pakanen, "International energy agency building demonstrating automated fault detection and diagnosis methods in real buildings.," tech. rep., Technical Research Centre of Finland, Laboratory of Heating and Ventilation, Espoo, Finland., 2001.

[30] S. Katipamula, R. G. Pratt, D. P. Chassin, Z. T. Taylor, K. Gowri, and M. R. Brambley, "Automated fault detection and diagnostics for outdoor-air ventilation systems and economizers: Methodology and results from field testing," *ASHRAE Transactions*, vol. 1, p. 105, 1999.

[31] L. K. Norford, J. A. Wright, R. A. Buswell, D. Luo, C. J. Klaassen, and A. Suby, "Demonstration of Fault Detection and Diagnosis Methods for Air-Handling Units," *HVAC&R Research*, vol. 8, pp. 41–71, Jan. 2002.

[32] T. A. Reddy and K. K. Andersen, "An Evaluation of Classical Steady-State Off-Line Linear Parameter Estimation Methods Applied to Chiller Performance Data," *HVAC&R Research*, vol. 8, pp. 101–124, Jan. 2002.

[33] T. A. Reddy, D. Niebur, K. K. Andersen, P. P. Pericolo, and G. Cabrera, "Evaluation of the Suitability of Different Chiller Performance Models for On-Line Training Applied to Automated Fault Detection and Diagnosis (RP-1139)," *HVAC&R Research*, vol. 9, pp. 385–

414, Oct. 2003.

[34] M. P. Gómez-Carracedo, J. M. Andrade, P. López-Mahía, S. Muniategui, and D. Prada, "A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets," *Chemometrics and Intelligent Laboratory Systems*, vol. 134, pp. 23–33, May 2014.

[35] Y. Ding and A. Ross, "A comparison of imputation methods for handling missing scores in biometric fusion," *Pattern Recognition*, vol. 45, pp. 919–933, Mar. 2012.

[36] F. O. de França, G. P. Coelho, and F. J. Von Zuben, "Predicting missing values with biclustering: A coherence-based approach," *Pattern Recognition*, vol. 46, pp. 1255–1266, May 2013.

[37] G. Olsson, M. Nielsen, Z. Yuan, A. Lynggaard-Jensen, and J.-P. Steyer, "Instrumentation, Control and Automation in Wastewater Systems," Scientific and Technical Report 15, May 2005.

[38] Z. Zhang, "Missing data exploration: highlighting graphical presentation of missing pattern," *Annals of Translational Medicine*, vol. 3, no. 22, 2015.

[39] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data," *Theoretical and Applied Climatology*, vol. 112, pp. 143–167, July 2012.

[40] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, pp. 2895–2907, June 2004.

[41] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, "Review: A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, pp. 1087–1091, Oct. 2006.

[42] A. Plaia and A. L. Bondí, "Single imputation method of missing values in environmental pollution data sets," *Atmospheric Environment*, vol. 40, pp. 7316–7330, Dec. 2006.

[43] R. R. Andridge and R. J. A. Little, "A Review of Hot Deck Imputation for Survey Non-response," *International statistical review = Revue internationale de statistique*, vol. 78, pp. 40–64, Apr. 2010.

[44] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques, Third Edition*. Burlington, MA: Morgan Kaufmann, 3 edition ed., July 2011.

[45] E. Frank and R. R. Bouckaert, "Conditional Density Estimation with Class Probability Estimators," in *Advances in Machine Learning* (Z.-H. Zhou and T. Washio, eds.), no. 5828 in Lecture Notes in Computer Science, pp. 65–81, Springer Berlin Heidelberg, 2009.

[46] L. Torgo and J. Gama, "Search-based Class Discretization," in *In: Proceedings of the Ninth European Conference on Machine Learning*, pp. 266–273, Springer Verlag, 1997.

[47] R. J. A. Little, "Regression with Missing X's: A Review," *Journal of the American Statistical Association*, vol. 87, pp. 1227–1237, Dec. 1992.

[48] A. P. Robinson and J. D. Hamann, "Imputation and Interpolation," in *Forest Analytics with R*, Use R, pp. 117–151, Springer New York, 2011. DOI: 10.1007/978-1-4419-7762-5_4.

[49] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *Journal of Systems and Software*, vol. 85, pp. 2541–2552, Nov. 2012.

[50] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 617–621, Oct. 1979.

[51] S. Katipamula and M. R. Brambley, "Review Article: Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems - A Review, Part I," *HVAC&R Research*, vol. 11, no. 1, pp. 3–25, 2005.

[52] N. Branisavljević, Z. Kapelan, and P. Dušan, "Improved real-time data anomaly detection

using context classification," *Journal of Hydroinformatics*, vol. 13, p. 307, July 2011.

[53] G. Tolle, J. Polastre, R. Szewczyk, D. Culler, N. Turner, K. Tu, S. Burgess, T. Dawson, P. Buonadonna, D. Gay, and W. Hong, "A Macroscope in the Redwoods," in *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems*, (New York, NY, USA), pp. 51–63, 2005.

[54] A. B. Sharma, L. Golubchik, and R. Govindan, "Sensor Faults: Detection Methods and Prevalence in Real-world Datasets," *ACM Trans. Sen. Netw.*, vol. 6, pp. 23:1–23:39, June 2010.

[55] A. Khan and K. Hornbæk, "Big Data from the Built Environment," in *Proceedings of the 2Nd International Workshop on Research in the Large*, LARGE '11, (New York, NY, USA), pp. 29–32, ACM, 2011.

[56] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava, "Sensor Network Data Fault Types," *ACM Trans. Sen. Netw.*, vol. 5, pp. 25:1–25:29, June 2009.

[57] S. Wang and J. Cui, "Sensor-fault detection, diagnosis and estimation for centrifugal chiller systems using principal-component analysis method," *Applied Energy*, vol. 82, pp. 197–213, Nov. 2005.

[58] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, and M. Welsh, "Fidelity and Yield in a Volcano Monitoring Sensor Network," in *Proceedings of the 7th Symposium on Operating Systems Design and Implementation*, (Berkeley, CA, USA), pp. 381–396, 2006.

[59] N. Ramanathan, L. Balzano, M. Burt, D. Estrin, T. Harmon, C. Harvey, J. Jay, E. Kohler, S. Rothenberg, and M.Srivastava, "Rapid Deployment with Confidence:Calibration and Fault Detection in Environmental Sensor Networks," *Center for Embedded Network Sensing*, Jan. 2006.

[60] N. Ramanathan, T. Schoellhammer, D. Estrin, M. Hansen, T. Harmon, E. Kohler, and M. Srivastava, "The Final Frontier: Embedding Networked Sensors in the Soil," *Center for Embedded Network Sensing*, Jan. 2006.

[61] V. Bychkovskiy, S. Megerian, D. Estrin, and M. Potkonjak, "A collaborative approach to in-place sensor calibration," in *Proceedings of the 2Nd International Conference on Information Processing in Sensor Networks*, IPSN'03, (Berlin, Heidelberg), pp. 301–316, Springer-Verlag, 2003.

[62] L. Balzano and R. Nowak, "Blind calibration of sensor networks," in *Proceedings of the 6th International Conference on Information Processing in Sensor Networks*, IPSN '07, (New York, NY, USA), pp. 79–88, ACM, 2007.

[63] V. Baljak, T. Kenji, and S. Honiden, "Faults in Sensory Readings: Classification and Model Learning," *Sensors & Transducers*, Jan. 2013.

[64] V. Baljak, K. Tei, and S. Honiden, "Classification of Faults in Sensor Readings with Statistical Pattern Recognition," pp. 270–276, Aug. 2012.

[65] V. Baljak, K. Tei, and S. Honiden, "Fault classification and model learning from sensory Readings - Framework for fault tolerance in wireless sensor networks," in *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 408–413, Apr. 2013.

[66] Chris Chatfield, *Time-Series Forecasting*. Boca Raton: Chapman and Hall/CRC, 1 edition ed., Oct. 2000.

[67] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. Hoboken, N.J: Wiley, 4 edition ed., June 2008.

[68] S. Hussain, M. Mokhtar, and J. Howe, "Sensor Failure Detection, Identification, and Accommodation Using Fully Connected Cascade Neural Network," *IEEE Transactions on*

Industrial Electronics, vol. 62, pp. 1683–1692, Mar. 2015.

[69] Z. Du, B. Fan, J. Chi, and X. Jin, "Sensor fault detection and its efficiency analysis in air handling unit using the combined neural networks," *Energy and Buildings*, vol. 72, pp. 157–166, Apr. 2014.

[70] G. Jäger, S. Zug, T. Brade, A. Dietrich, C. Steup, C. Moewes, and A.-M. Cretu, "Assessing neural networks for sensor fault detection," in *2014 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pp. 70–75, May 2014.

[71] Portland Energy Conservation Incorporation and Battelle Northwest Division, "Methods for automated and continuous commissioning of building systems," Tech. Rep. ARTI-21CR/610-30040-01, Apr. 2003.

[72] S. Katipamula, M. R. Brambley, and L. Luskay, "Automated Proactive Techniques for Commissioning Air-Handling Units," *Journal of Solar Energy Engineering*, vol. 125, pp. 282–291, Aug. 2003.

[73] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods," *Computers & Chemical Engineering*, vol. 27, pp. 293–311, Mar. 2003.

[74] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies," *Computers & Chemical Engineering*, vol. 27, pp. 313–326, Mar. 2003.

[75] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis: Part III: Process history based methods," *Computers & Chemical Engineering*, vol. 27, pp. 327–346, Mar. 2003.

[76] S. Bendapudi and J. E. Braun, "A review of literature on dynamic models of vapor compression equipment," Tech. Rep. 4036-5, Ray Herrick Laboratories, Purdue University, 2002.

[77] J. Wagner and R. Shoureshi, "Failure Detection Diagnostics for Thermofluid Systems," *Journal of Dynamic Systems, Measurement, and Control*, vol. 114, pp. 699–706, Dec. 1992.

[78] P. Haves, T. Salsbury, and J. A. Wright, "Condition Monitoring in HVAC Subsystems using First Principles Models," *ASHRAE Transactions*, vol. 102, no. Pt. 1, 1996.

[79] T. I. Salsbury and R. C. Diamond, "Fault detection in HVAC systems using model-based feedforward control," *Energy and Buildings*, vol. 33, pp. 403–415, Apr. 2001.

[80] K. Oosterhuis, "Simply complex, toward a new kind of building," *Frontiers of Architectural Research*, vol. 1, no. 4, pp. 411 – 420, 2012.

[81] E. Azar and C. Menassa, "A conceptual framework to energy estimation in buildings using agent based modeling," in *Simulation Conference (WSC), Proceedings of the 2010 Winter*, pp. 3145–3156, Dec 2010.

[82] E. Azar and C. C. Menassa, "Agent-based modeling of occupants and their impact on energy use in commercial buildings," *Journal of Computing in Civil Engineering*, vol. 26, no. 4, pp. 506–518, 2011.

[83] T. Jensen, G. Holtz, C. Baedeker, and É. J. Chappin, "Energy-efficiency impacts of an air-quality feedback device in residential buildings: an agent-based modeling assessment," *Energy and Buildings*, 2016.

[84] V. Grimm, E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H.-H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis, "Pattern-oriented modeling of agent-based complex systems: lessons from ecology," *science*, vol. 310, no. 5750, pp. 987–991, 2005.

[85] C. Ghiaus, "Fault diagnosis of air conditioning systems based on qualitative bond graph," *Energy and Buildings*, vol. 30, pp. 221–232, Aug. 1999.

[86] Arthur Dexter and Jouko Pakanen, "Demonstrating Automated Fault Detection and Di-

agnosis Methods in Real Buildings," Tech. Rep. Annex 34, International Energy Agency, Energy conservation in buildings and community systems, Technical Research Centre of Finland, Laboratory of Heating and Ventilation, Espoo, Finland., 2001.

[87] M. Brambley, R. Pratt, D. Chassin, S. Katipamula, and D. Hatley, "Diagnostics for outdoor air ventilation and economizers," *ASHRAE journal*, vol. 40, no. 10, p. 49, 1998.

[88] J. M. House, H. Vaezi-Nejad, and J. M. Whitcomb, "An expert rule set for fault detection in air-handling units," *Transactions-American Society of Heating Refrigerating and Air Conditioning Engineers*, vol. 107, no. 1, pp. 858–874, 2001.

[89] T. M. Rossi and J. E. Braun, "A statistical, rule-based fault detection and diagnostic method for vapor compression air conditioners," *HVAC&R Research*, vol. 3, no. 1, pp. 19–37, 1997.

[90] M. S. Breuker and J. E. Braun, "Common faults and their impacts for rooftop air conditioners," *HVAC&R Research*, vol. 4, no. 3, pp. 303–318, 1998.

[91] H. Grimmelius, J. K. Woud, and G. Been, "On-line failure diagnosis for compression refrigeration plants," *International Journal of Refrigeration*, vol. 18, no. 1, pp. 31 – 41, 1995.

[92] M. Stylianou and D. Nikanpour, "Performance monitoring, fault detection, and diagnosis of reciprocating chillers," vol. 102, pp. 615–627, ASHRAE Transactions, 1996.

[93] J. Gordon and K. C. Ng, "Predictive and diagnostic aspects of a universal thermodynamic model for chillers," *International Journal of Heat and Mass Transfer*, vol. 38, no. 5, pp. 807 – 818, 1995.

[94] M. Stylianou, "Application of classification functions to chiller fault detection and diagnosis," vol. 103, pp. 645–648, ASHRAE Transactions, 1997.

[95] G. M. Kaler, "Embedded Expert System Development for Monitoring Packaged HVAC Equipment," in *ASHRAE Transactions (American Society of Heating, Refrigerating and Air- Conditioning Engineers)*, vol. 96, p. 733, 1990.

[96] S. Kaldorf and P. Gruber, "Practical experiences from developing and implementing an expert system diagnostic tool/discussion," *ASHRAE Transactions*, vol. 108, p. 826, 2002.

[97] S. A. Patel and A. K. Kamrani, "Intelligent decision support system for diagnosis and maintenance of automated systems," *Computers & Industrial Engineering*, vol. 30, no. 2, pp. 297 – 319, 1996.

[98] J. De Kleer and J. S. Brown, "A qualitative physics based on confluences," *Artificial intelligence*, vol. 24, no. 1, pp. 7–83, 1984.

[99] H. Kay, "Sqsim: a simulator for imprecise {ODE} models," *Computers & Chemical Engineering*, vol. 23, no. 1, pp. 27 – 46, 1998.

[100] R. Zmeureanu, "Prediction of the cop of existing rooftop units using artificial neural networks and minimum number of sensors," *Energy*, vol. 27, no. 9, pp. 889 – 904, 2002.

[101] C. Fan, F. Xiao, and C. Yan, "A framework for knowledge discovery in massive building automation data and its application in building diagnostics," *Automation in Construction*, vol. 50, pp. 81–90, Feb. 2015.

[102] P. L. Riemer, J. W. Mitchell, and W. A. Beckman, "The use of time series analysis in fault detection and diagnosis methodologies," *ASHRAE Transactions*, vol. 108, p. 384, 2002.

[103] K. K. Andersen and T. A. Reddy, "The Error in Variables (EIV) Regression Approach as a Means of Identifying Unbiased Physical Parameter Estimates: Application to Chiller Performance Data," *HVAC&R Research*, vol. 8, pp. 295–309, July 2002.

[104] A. Capozzoli, F. Lauro, and I. Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Systems with Applications*, vol. 42, pp. 4324–4338, June 2015.

[105] W.-Y. Lee, C. Park, and G. E. Kelly, "Fault detection in an air-handling unit using resid-

ual and recursive parameter identification methods," *Transactions-American Society Of Heating Refrigerating And Air Conditioning Engineers*, vol. 102, pp. 528–539, 1996.

[106] W.-Y. Lee, J. M. House, C. Park, and G. E. Kelly, "Fault diagnosis of an air-handling unit using artificial neural networks," *Transactions - American society of Heating Refrigerating and Air Conditioning Engineers*, vol. 102, pp. 540–549, 1996.

[107] X. Li, J. Visier, and H. Vaezi-Nejad, "A neural network prototype for fault detection and diagnosis of heating systems," *Ashrae Transactions*, vol. 103, no. 1, pp. 634–644, 1997.

[108] T. Reddy, "Application of dynamic building inverse models to three occupied residences monitored non-intrusively," *Proceedings of the Thermal Performance of the Exterior Envelope of Buildings IV*, 1989.

[109] G. Guyon and E. Palomo, "Validation of two french building energy programs, part 2: Parameter estimation method applied to empirical validation," *ASHRAE Transactions*, vol. 105, p. 709, 1999.

[110] Y. Jia and T. A. Reddy, "Characteristic Physical Parameter Approach to Modeling Chillers Suitable for Fault Detection, Diagnosis, and Evaluation," *Journal of Solar Energy Engineering*, vol. 125, pp. 258–265, Aug. 2003.

[111] D. Pietruschka, A. Dalibard, I. B. Hassine, H. Focke, F. Judex, A. Preisler, M. Helm, P. Ohnewein, and A. Frein, "Report for self-detection on monitoring procedure," Tech. Rep. IEA Task 48, International Energy Agency, Oct. 2015.

[112] V. Figueiredo, F. Rodrigues, Z. Vale, and J. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Transactions on Power Systems*, vol. 20, pp. 596–602, May 2005.

[113] M. Domínguez, J. J. Fuertes, S. Alonso, M. A. Prada, A. Morán, and P. Barrientos, "Power monitoring system for university buildings: Architecture and advanced analysis tools," *Energy and Buildings*, vol. 59, pp. 152–160, Apr. 2013.

[114] N. Djuric and V. Novakovic, "Identifying important variables of energy use in low energy office building by using multivariate analysis," *Energy and Buildings*, vol. 45, pp. 91–98, Feb. 2012.

[115] J. E. Seem, "Pattern recognition algorithm for determining days of the week with similar energy consumption profiles," *Energy and Buildings*, vol. 37, pp. 127–139, Feb. 2005.

[116] J. E. Seem, "Using intelligent data analysis to detect abnormal energy consumption in buildings," *Energy and Buildings*, vol. 39, pp. 52–58, Jan. 2007.

[117] A. Kusiak and Z. Song, "Clustering-Based Performance Optimization of the Boiler -Turbine System," *IEEE Transactions on Energy Conversion*, vol. 23, pp. 651–658, June 2008.

[118] A. R. Florita, L. J. Brackney, T. P. Otanicar, and J. Robertson, "Classification of Commercial Building Electrical Demand Profiles for Energy Storage Applications," *Journal of Solar Energy Engineering*, vol. 135, pp. 031020–031020, June 2013.

[119] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom, "Visually mining and monitoring massive time series," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, (New York, NY, USA), pp. 460–469, ACM, 2004.

[120] E. Keogh and S. Kasetty, "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration," *Data Mining and Knowledge Discovery*, vol. 7, pp. 349–371, Oct. 2003.

[121] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," *Data mining in time series databases*, vol. 57, pp. 1–22, 2004.

[122] K. Z. Haigh, W. Foslien, and V. Guralnik, "Visual query language: Finding patterns in and relationships among time series data," in *Seventh Workshop on Mining Scientific and*

Engineering Datasets, vol. 24, 2004.

[123] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11, ACM, 2003.

[124] J. Zakaria, S. Rotschafer, A. Mueen, K. Razak, and E. J. Keogh, "Mining massive archives of mice sounds with symbolized representations.," in *SDM*, pp. 588–599, SIAM, 2012.

[125] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 262–270, ACM, 2012.

[126] S. Laxman and P. S. Sastry, "A survey of temporal data mining," *Sadhana*, vol. 31, no. 2, pp. 173–198, 2006.

[127] V. Kavitha and M. Punithavalli, "Clustering time series data stream - a literature survey," *International Journal of Computer Science and Information Security*, vol. 8, no. 1, pp. 289–294, 2010.

[128] T. W. Liao, "Clustering of time series data - a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[129] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering - a decade review," *Information Systems*, vol. 53, pp. 16 – 38, 2015.

[130] S. Rani and G. Sikka, "Recent techniques of clustering of time series data: a survey," *International Journal of Computer Applications*, vol. 52, no. 15, 2012.

[131] C. M. Antunes and A. L. Oliveira, "Temporal data mining: An overview," in *KDD workshop on temporal data mining*, vol. 1, p. 13, 2001.

[132] M. Chiş, S. Banerjee, and A. E. Hassanien, "Clustering time series data: an evolutionary approach," in *Foundations of Computational, IntelligenceVolume 6*, pp. 193–207, Springer, 2009.

[133] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, pp. 107–144, Apr. 2007.

[134] K.-P. Chan and A. W.-C. Fu, "Efficient time series matching by wavelets," in *Data Engineering, 1999. Proceedings., 15th International Conference on*, pp. 126–133, IEEE, 1999.

[135] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, pp. 358–386, May 2004.

[136] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: a survey and empirical demonstration," *Data Mining and knowledge discovery*, vol. 7, no. 4, pp. 349–371, 2003.

[137] C. A. Ratanamahatana and E. Keogh, "Making time-series classification more accurate using learned constraints," SIAM, 2004.

[138] A. Nanopoulos, R. Alcock, and Y. Manolopoulos, "Feature-based classification of time-series data," *International Journal of Computer Research*, vol. 10, no. 3, pp. 49–61, 2001.

[139] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data mining and knowledge Discovery*, vol. 13, no. 3, pp. 335–364, 2006.

[140] R. Agrawal, C. Faloutsos, and A. Swami, *Efficient similarity search in sequence databases*. Springer, 1993.

[141] K. Deng, A. W. Moore, and M. C. Nechyba, "Learning to recognize time series: Combining arma models with memory-based learning," in *Computational Intelligence in Robotics and Automation, 1997. CIRA'97., Proceedings., 1997 IEEE International Symposium on*, pp. 246–251, IEEE, 1997.

[142] X. Ge and P. Smyth, "Deformable markov model templates for time-series pattern matching," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 81–90, ACM, 2000.

[143] T. Kinnunen, I. Sidoroff, M. Tuononen, and P. Fränti, "Comparison of clustering methods: A case study of text-independent speaker modeling," *Pattern Recognition Letters*, vol. 32, no. 13, pp. 1604–1617, 2011.

[144] S. J. Fodeh, W. F. Punch, and P.-N. Tan, "Combining statistics and semantics via ensemble model for document clustering," in *Proceedings of the 2009 ACM symposium on Applied Computing*, pp. 1446–1450, ACM, 2009.

[145] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 89–98, ACM, 2006.

[146] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[147] R. Caruana, M. Elhawary, N. Nguyen, and C. Smith, "Meta clustering," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pp. 107–118, IEEE, 2006.

[148] C. Goutte, L. K. Hansen, M. G. Liptrot, and E. Rostrup, "Feature-space clustering for fMRI meta-analysis," *Human Brain Mapping*, vol. 13, pp. 165–183, July 2001.

[149] P. Hill, T. & Lewicki, *Statistics Methods and Applications Book*. StatSoft, Inc., 2013.

[150] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, pp. 411–423, Jan. 2001.

[151] D. J. Ketchen and C. L. Shook, "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique," *Strategic Management Journal*, vol. 17, pp. 441–458, June 1996.

[152] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[153] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, pp. 224–227, Apr. 1979.

[154] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, pp. 1–27, Jan. 1974.

[155] T. Selke and A. Preisler, "Energybase-sunny office future," in *9th International Symposium for Solar Thermal Energy Utilization (SOLAR 2008)*, 2008.

[156] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, Springer, 2007.

[157] G. Zucker, U. Habib, M. Blöchle, F. Judex, and T. Leber, "Sanitation and Analysis of Operation Data in Energy Systems," *Energies*, vol. 8, pp. 12776–12794, Nov. 2015.

[158] G. Zucker, U. Habib, M. Blöchle, A. Wendt, S. Schaat, and L. C. Siafara, "Building energy management and data analytics," in *2015 International Symposium on Smart Electric Distribution Systems and Technologies (EDST)*, pp. 462–467, Sept. 2015.

[159] G. Zucker, J. Malinao, U. Habib, T. Leber, A. Preisler, and F. Judex, "Improving energy efficiency of buildings using data mining technologies," in *2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)*, pp. 2664–2669, June 2014.

[160] U. Habib, G. Zucker, M. Blochle, F. Judex, and J. Haase, "Outliers detection method using clustering in buildings data," in *Industrial Electronics Society, IECON 2015 - 41st Annual Conference of the IEEE*, pp. 000694–000700, Nov 2015.

[161] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[162] O. Elnokity, I. I. Mahmoud, M. K. Refai, and H. M. Farahat, "ANN based sensor faults detection, isolation, and reading estimates - SFDIRE: Applied in a nuclear process," *Annals of Nuclear Energy*, vol. 49, pp. 131 – 142, 2012.

[163] M. Transell and S. Carl, "The use of bioinformatics techniques for time-series motif-matching: A case study," in *The Sixth International Conference on Advanced Engineering Computing and Applications in Sciences*, pp. 114–117, IARIA, 2012.

[164] A. McGovern, D. Rosendahl, R. Brown, and K. Droegemeier, "Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 232–258, 2011.

[165] M. Junker, R. Hoch, and A. Dengel, "On the evaluation of document analysis components by recall, precision, and accuracy," in *Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999. ICDAR '99*, pp. 713–716, Sept. 1999.

[166] M. Hossin, M. Sulaiman, A. Mustapha, N. Mustapha, and R. Rahmat, "A hybrid evaluation metric for optimizing classifier," in *Data Mining and Optimization (DMO), 2011 3rd Conference on*, pp. 165–170, June 2011.

[167] U. Habib and G. Zucker, "Finding the different patterns in buildings data using bag of words representation with clustering," in *2015 13th International Conference on Frontiers of Information Technology (FIT)*, pp. 303–308, Dec 2015.

[168] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, pp. 411–423, Jan. 2001.

[169] S. Saraçli, N. Doğan, and İ. Doğan, "Comparison of hierarchical cluster analysis methods by cophenetic correlation," *Journal of Inequalities and Applications*, vol. 2013, pp. 1–8, Apr. 2013.

[170] Ministry of Planning and Development Government of Pakistan, "Pakistan Integrated Energy Model (Pak-IEM), Final Report Volume II Policy Analysis Report ," Tech. Rep. ADB TA-4982 PAK, Islamabad, Pakistan, December 2010.

Curriculum Vitae

Usman Habib

Längenfeldgasse 22/16,
1120, Vienna, Austria.
Mobile: +43-6603442231
Email:usmansagar@gmail.com

Objective:-

I want to be a part of a dynamic organization, to get professional experience while utilizing my potential for achieving excellence and applying all that I have learned from experience and academics.

Education

Degree	Institute	Year	CGPA/ Score
PhD (In Progress)	Technische Universität Wien, Vienna, Austria	2013-Till Date	In Progress
MS In Engineering (Telematics - Communication Networks and Networked Services) Specialization in Information Security.	Norwegian University of Science and Technology (NTNU), Norway	2008-2010	Completed
BS in Computer Science	COMSATS Institute of Information Technology Abbottabad, Pakistan	2002-2006	3.4/4 (83%) (Gold Medal)

Experience

Title	Organization	Started	To	Experience
Research Associate	Austrian Institute of Technology (AIT), Vienna, Austria	June 2013	Till Date	Working on research project funded by the Austrian Funding Agency in the funding programme e!MISSION within the project "extrACT".
Assistant Professor	COMSATS Institute of Information Technology Abbottabad.	July 2012	May 2013	Taught courses of Introduction to Programming, Data Structures, and Object Oriented Programming, Network Operating Systems, Compiler Construction.
Lecturer	COMSATS Institute of Information Technology Abbottabad.	August 2006	June 2012	Taught courses of Introduction to Programming, Data Structures, and Object Oriented Programming, Network Operating Systems, Compiler Construction.
Thesis Research	Telenor R&I Norway/ Ubisafe AS Norway	September 2009	June 2010.	The thesis has been carried out in collaboration with Telenor R&I and Ubisafe AS Norway
Programmer	IT-SOL U.K	January 2006	December 2006	Free launce programmer(part time job)
Volunteer (IT Specialist)	Pakistan Army (14th Para Brigade Shinkiari)	November 2005	December 2005	Maintain data base of earth quake effecties. Established a network.
Volunteer (IT Specialist)	UNICEF & UNDP	December 2005	January 2006	Made a software project that helps them to maintain the data of children that were earthquake effecties

Book Published:

- **Usman Habib**, "Secure Mobile Authentication for Linux Workstation log on", LAMBERT Academic Publishing, 2011, ISBN: 978-3-8454-0289-5

Journal Papers:

- G. Zucker, **U. Habib**, M. Blöchle, F. Judex, and T. Leber, "Sanitation and Analysis of Operation Data in Energy Systems," Energies, vol. 8, no. 11, pp. 12776–12794, Nov. 2015. (IF: 2.072)
- **U. Habib**, I. Jørstad, D. V. Thanh, K. Hayat, and I. A. Khan, "A Framework for Secure Linux Based Authentication in Enterprises via Mobile Phone," J. Basic Appl. Sci. Res., vol. 1, no. 1(12), pp. 3058–3066, 2011.
- S. Khan and **U. Habib**, "Procedural Justice & Organizational Performance," Abasyn J. Soc. Sci., vol. 4, no. 7, pp. 36–51, 2011.

Conference Papers:

- **U. Habib** and G. Zucker, "Finding the Different Patterns in Buildings Data Using Bag of Words Representation with Clustering," in 2015 13th International Conference on Frontiers of Information Technology (FIT), 2015, pp. 303-308.
- **U. Habib**, G. Zucker, M. Blochle, F. Judex, and J. Haase, "Outliers detection method using clustering in buildings data," in IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society, 2015, pp. 000694-000700.
- G. Zucker, **U. Habib**, M. Blochle, A. Wendt, S. Schaat, and L. C. Siafara, "Building energy management and data analytics," in 2015 International Symposium on Smart Electric Distribution Systems and Technologies (EDST), 2015, pp. 462–467.
- G. Zucker, J. Malinao, **U.Habib**, T. Leber, A. Preisler, and F. Judex, "Improving energy efficiency of buildings using data mining technologies," in 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE), 2014, (PP) 2664–2669.
- M. A. Tawab Khalil, P. D. D. Dominic, H. Kazemian, and **U. Habib**, "A Study to Examine If Integration of Technology Acceptance Model's (TAM) Features Help in Building a Hybrid Digital Business Ecosystem Framework for Small and Medium Enterprises (SMEs)," in Frontiers of Information Technology (FIT), 2011, 2011, pp. 161–166.

BS Projects Supervised:

- Speaking Dumb on a Telephone.
- Conference management system.
- Login using mobile.

Master's thesis: -"Secure mobile authentication for Linux workstation log on"

The Master thesis proposed how workstation identity management can be made more user-friendly and secure by using the mobile phone in the Linux workstation logon process. The thesis has been carried out in collaboration with Telenor R&I Norway and Ubisafe AS Norway. The thesis was supervised by Dr. Do Van Thanh (Senior Scientist Telenor R&I Norway) and the co-supervisor was Dr. Ivar Jørstad (CEO Ubisafe AS)

(Bachelors) Final Year Project: - ***Iris Detection and Recognition System***
Human Iris Detection and recognition system is developed to uniquely identify a person with his iris patterns. As every human has unique iris patterns which will help to recognize them.

Other Projects

- Website for Adan Consultants (www.adanconsultant.com).
- Data Encryption and Decryption using RSA algorithm for key generation and steganography.
- Office Management system for IT-SOL software house.
- A system for UNICEF for maintaining children information during earthquake operation.
- A system for UNDP for maintaining item records during earthquake relief operation.

Awards/Achievements

- Higher Education Commission (HEC) of Pakistan Scholarship for PhD at Technical University of Vienna, Austria.
- Higher Education Commission (HEC) of Pakistan Scholarship for Masters at NTNU, Norway.
- Held **sponsorships** from COMSATS from 2002 to 2006.
- Holding first position for 2006 batch.
- Holding **Campus gold medal** for 2006 batch.
- University Cricket Team Captain.
- Worked In a group of IT members in Shinkiari with Army to establish and maintain data of earthquake disaster.
- Worked in organizer committee for ICT's Conference, 2006.
- Worked in organizer Committee for FIT 2006(Frontiers of Information Technology) international conference held on 20th and 21st December, 2006 at Marriott Hotel Islamabad.
- Worked in organizing committee for "VisionICT" held in COMSATS Abbottabad.
- Worked in organizing committee for Convocation 2007 Abbottabad.
- Worked in organizing committee for CWRC (CIIT workshop on research and computing) held on 27th October, 2007 at Abbottabad.
- Worked in organizing committee for FIT 2007 (Frontiers of Information Technology) international conference held on 17th and 18st December, 2007 at Marriott Hotel Islamabad.
- Worked in organizing committee for FIT 2010 (Frontiers of Information Technology) 8th International conference held from 21st December to 23rd December, 2010 at Islamabad.
- Worked in organizing committee for FIT 2011 (Frontiers of Information Technology) 9th International conference held from 19th December to 21st December, 2011 at Islamabad.
- Worked in organizing committee for FIT 2012 (Frontiers of Information Technology) 10th International conference held from 17th December to 19th December, 2012 at Islamabad.