Development and implementation of a framework to aid the transition to proactive maintenance approaches on air handling units in the industrial setting

Thesis presented by

Michael Ahern BEng (Hons)

119227513

for the degree of

Doctor of Philosophy

University College Cork

School of Engineering

Head of School: Prof. Jorge Oliveira

Supervisors: Dr. Ken Bruton, Dr. Dominic O'Sullivan

2023

Research supported by Science Foundation Ireland through the MaREI Centre

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

AFDD	Automated fault detection and diagnostics
AHU	Air Handling Unit
AIR	Adjacent Information Recovery
ANN	Articial Neural Network
APAR	AHU Performance Assessment Rules
ARM	Association Rule Mining
ARX	Autoregressive
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CART	Classification and Regression Tree
CC-RF	Classifier Chains-Random Forest
CCx	Continuous Commissioning
CNN	Convolution Neural Network
CO_2	Carbon Dioxide
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CSV	Comma-Separated Values
CWGAN	Conditional Wasserstein Generative Adversarial Network
DBN	Diagnostic Bayesian Network

DENiM	Digital intelligence for collaborative ENergy management in Manufacturing
DL	Deep Learning
DST-map	Data Science Trajectories Map
DT	Decision Tree
DTO-DRNN	Deep Transition Output-Deep Recurrent Neural Networks
ECA	Energy Consumption Anomaly
EIA	Energy Information Administration
EnPI	Energy Performance Indicator
EPBD	Energy Performance of Buildings Directive
EU	European Union
FDD	Fault Detection and Diagnosis
GAN	Generative Adversarial Network
GHG	Greenhouse gas
GUI	Graphical User Interface
HMM	Hidden Markov Model
HRF	Hybrid Random Forest
HVAC	Heating, Ventilation and Air Conditioning
IDAIC	Industrial Data Analysis Improvement Cycle

IEA	International Energy Agency
IRENA	International Renewable Energy Agency
KDD	Knowledge Discovery in Databases
LSM	Least-Squares Method
LSTM-SVDD	Long Short-Term Memory-Support Vector Data Description
MBCx	Monitoring-based Commissioning
MCNN	Multiscale Convolution Neural Networks
MLP	Multi-Layer Perceptron
MPC	Model Predictive Control
MTU	Munster Technological University
NN	Neural Networks
OC-SVM	One-class Support Vector Machine
OCx	Ongoing Commissioning
PCA	Principal Component Analysis
RACNN	Rule-based method and Convolution Neural Network
RCx	Recommissioning
RDF	Resource Description Framework
Retro Cx	Retrocommissioning

RF	Random Forests
RNN	Recurrent Neural Network
SAE	Supervised Auto-Encoder
SVM	Support Vector Machine
SVR	Support Vector Regresssion
TARM	Temporal Association Rules Mining
TBKSFA	Three-way Data Based Kernel Slow Feature Analysis
TRNSYS	Transient System Simulation Software
UCC	University College Cork
WGAN	Wasserstein Generative Adversarial Network
XGBoost	Extreme Gradient Boosting

DECLARATION

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Michael Ahern   Date

Ken Bruton   Date

ABSTRACT

This research explores the transition to proactive maintenance of HVAC equipment, specifically AHUs in industrial facilities, to make progress towards climate goals. HVAC systems account for 14% of global energy consumption, with the potential to increase efficiency by 20% by addressing energy-wasting faults according to Roth et al. However, these faults are difficult to detect due to their compensating control logic. This research highlights the potential of digitalisation techniques, particularly AI, to identify and rectify these faults, which would contribute to an approximate 2.8% global efficiency improvement. Notably, the study focuses on the nuances of industrial facilities, which have received limited attention compared to other building types. The research identifies several gaps in existing literature, including the knowledge gap between proactive data analysis and reactive engineering mind-sets, the data gap between high-quality experimental datasets and poor-quality industrial datasets, the operational gap between known baselines in experimental studies and unknown baselines in industrial settings, and the practice-theory gap between data-driven approaches in the literature and rule-based approaches in commercial tools. To address the knowledge gap, this thesis presents the IDAIC framework, a domain knowledge integration-type adaptation of the CRISP-DM process model. The implementation of the framework in an industrial facility to curate a dataset and develop a data assessment decision tree has contributed towards closing the data gap. Additionally, the study proposes extensions to the APAR ruleset and a practical data-driven fault detection method to address the operational gap. The deployment of the IDAIC framework as a tool leverages the UML modelling language to address practical considerations and demonstrate the approach's flexibility. Therefore, the main research outputs include the IDAIC framework, an industrial AHU dataset, a data assessment decision tree, an extension to the APAR ruleset, and a proactive maintenance decision support tool. Notably, this research unveils a fault in which the outside air damper is stuck in the fully open position, leading to estimated annual savings of €60,000. These findings validate the effectiveness of a human-centric, domain expertise-integrated approach that is resilient to industrial challenges, contributing to sustainable energy efficiency improvements and the achievement of climate targets using the best available solutions.

ACKNOWLEDGEMENTS

I would like to thank Dr. Ken Bruton for the opportunity to embark on this research and his commitment to providing the very best of support and guidance to not only myself but all those that have been fortunate enough to have had Ken as a mentor. I always felt that Ken had my best interests at heart and would ensure that I developed the skills needed to become effective in both academia and industry. I would also like to express a special thanks to Dr. Dominic O'Sullivan in this regard for his valuable advice, knowledge and support throughout this research.

To my IERG colleagues and those past and present in the Irish Ashley Cummins building, I would like to thank you for the breaks, chats and consoling advice that can only be provided by those experiencing similar challenges. There are too many to mention individually but I would like to especially thank Adrita, Rosie, Alex, Liam, Suzanne, Sid, Talie and Akshay for broadening my cultural knowledge (especially regarding cuisine!) and inspiring some travel ideas for life post PhD. I'd also like to extend thanks to all the attendees of the morning coffee, especially Richie for sorting me out with keys and all things UCC that's off the research track.

I also couldn't have carried out such impactful research without the great access and support provided by DePuy Synthes and all those that found time in their busy schedules to help me. I would like to especially mention Eymard Gorman, Aidan Cloonan and Luke McAuliffe. Additionally, I appreciate the funding from MaREI through SFI and their continuous efforts to ensure that the structures are in place to provide a collaborative and supporting environment for researchers. Furthermore, I'd like to thank Andrew de Juan and Maxime Louvel for their efforts in developing the DENiM tool.

To my friends and past housemates in "38" Cathal, Conor and Lee, thanks for making the difficult COVID-19 era one with which I look back on with fond memories. I'd also like to thank Aaron, Brendan, Conor, Oisín and Ryan for the much-needed distractions.

Finally thanks to my Mam and Dad, Margaret and Dónal Ahern and to my brother Jack and sister Sarah for their help, and especially their patience with receiving the same status report on asking throughout this PhD – "It's Grand". I'll be forever grateful.

1 CHAPTER 1 - INTRODUCTION

1.1 Research Problem

It is our duty as custodians of the planet to ensure that the world is left in a state fit for future generations. It is the most fundamental existential challenge. Logically, it would appear to be a matter of maintaining the course and trusting traditional approaches successfully adopted by our ancestors to lead us to where we are today. However, the world is on an unsustainable path. Increasing global surface temperatures are adversely impacting our most basic needs such as water availability and food production [1]. Coupled with other global issues such as war and pandemics, the severity and complexity of our challenge quickly becomes apparent. No "silver-bullet" or "panacea" is readily available to cure the world of all its ailments, nor is a singular treatment at hand to manage its underlying health conditions or chronic diseases. When the issues faced are so diverse in nature, it is easy to become overwhelmed. For this reason, coordinated efforts are needed to contextualise the most significant issues we face and then address these issues using the best solutions at our disposal in order to reach a sustainable future.

In 2015, the United Nations (UN) released the 17 Sustainable Development Goals (SDGs) [2]. One of the key mechanisms to deliver on the SDGs is through better energy consumption management as energy consumption is the single largest contributor to climate change [3] and it affects many of the SDGs such as Goal 7 – Affordable and Clean Energy, Goal 12 – Responsible Consumption and Production, and Goal 13 – Climate Action. A new means of achieving better energy consumption management has been facilitated through the advent of digital technologies. The EU has proposed digitalisation as a key driver for transformation that will keep Europe both sustainable and competitive [4]. However, while the digital transformation of Europe's economy accelerated due to the COVID-19 crisis, firms put more complex digitalisation processes on hold [5]. This means that valuable data is being recorded to various degrees of quantity and quality, but it is not being fully leveraged in terms of assisting energy consumption reduction. Therefore, the main motivation for this research is to elicit the untapped potential of data to deliver the twin ecological and digital transitions [6]. Further research is needed to find a means of replicating the advanced methods proposed in the literature in the real-world.

According to the IEA World Energy Outlook 2022, industry (167 EJ) and buildings (132 EJ) account for a combined 68% of total final energy consumption (439 EJ) in 2021 [7]. A system that is common to almost all indoor facilities in these sectors and accounts for 14% of energy use worldwide [8], is heating, ventilation, and air conditioning (HVAC) systems. There is high potential for energy savings in HVAC systems as their self-correcting nature compensates for faulty components or operation, facilitating the occurrence of unnoticed faults that may account for up to 20% of the energy consumed by these systems [9]. Monetarily, Mills [10] estimated that common faults such as duct leakage, dampers not working optimally, and valve leakage cost 2.9, 0.5, and 0.1 $billion/year, respectively, in commercial buildings, based on the findings by Roth et al. [9]. HVAC systems therefore need to be well maintained, but maintenance costs, time for repair, and replacement of components that have not reached the end of their useful life must be balanced with the potential energy savings.

Therefore, proactive maintenance of HVAC systems is an important area of research as it is a low cost means of contributing to climate targets while also including the added benefits of reduced downtime, optimised use of spare parts, and improved asset reliability [11]. In this approach, data-driven analysis is used to identify the early signs of degradation, which enables maintenance activity to be scheduled at the optimal time [12]. This results in the maintenance activity being carried out only when it is necessary, thus minimising both energy waste and the probability of catastrophic failure which leads to downtime. Most of the research in the literature is focused on commercial or academic buildings, with few studies highlighting the nuances of the challenges in industrial facilities. These include highly regulated facilities, data acquisition difficulties, capital expenditure constraints, and system complexity. ***Therefore, developing a proactive maintenance solution for air handling units (AHUs), a main component in HVAC systems, in industrial facilities is a novel area of research that contributes to climate targets, reduces maintenance costs, reduces downtime, and improves operational efficiency.***

1.2 Research Question

The key issues that contribute to the research problem of developing a proactive maintenance strategy for AHUs are the knowledge gap, the data gap, the operational gap, and the practice-theory gap.

The data-driven approach requires skills from the field of data science. Data scientists possess strong analytical ability but generally lack fundamental knowledge of the physical operation of the equipment. Engineers possess this contextual knowledge, but they are generally unaware of the capabilities of data-driven approaches. To direct the power of analytics in a direction that provides value, this ***knowledge gap*** must be bridged with a solution that combines the advantages of both disciplines. This is a challenging objective considering there is a lack of a process to follow in the literature for merging these two fields that have their own philosophies, concepts, and methodologies.

Therefore, most of the data-driven analysis in the literature has been conducted by researchers with strong algorithmic ability on carefully curated simulated or laboratory-based datasets. While these datasets provide a good benchmark to compare model performance, they differ from real-world data in a few key areas. First, experimental and simulated datasets are well labelled with key information such as fault type, while real-world datasets with confirmed and reliable reference data or verified ground truth are extremely rare [13]. Second, the quality of experimental datasets is very high, while real-world datasets are often incomplete and inaccurate. Third, IT and data integration represent one of the largest barriers to scale [14] as real-world datasets are difficult to prepare when considering bespoke naming conventions and data collection methods. This creates a ***data gap*** that must be filled to increase the reproducibility and applicability of techniques published in literature.

Similarly, the operation of real-world AHUs in the industrial setting is inherently different from the AHUs analysed in the literature. For example, experimental studies input fault scenarios for a known length of time at a known severity or intensity [15]. In practice, AHU performance degrades over time, so it is difficult to determine the exact point at which a fault occurs, the type of fault that is occurring, and the severity of the fault. This causes difficulty in determining when maintenance or commissioning in the case of

HVAC should occur. This is a significant issue as Claridge et al. [16] found that an increase of 12.1% in overall building heating and cooling is possible two years after commissioning. As it is often unknown which faults are active in a building, training data may be mislabelled, which would ultimately lead to incorrect predictions from a data-driven approach [8]. Given the difficulty in understanding the current health state of an AHU in the industrial setting, a readiness indicator is needed for the point at which data-driven approaches would potentially begin to provide value. Coupled with the lack of domain-guided analysis in this field [17], an ***operational gap*** exists between the current unknown operation of AHUs in practical settings and the optimal starting point for data-driven applications.

On considering these three gaps, there is a logical final gap between commercial tools and articles in the literature. In a review conducted by Zhao et al. [18], the authors found that 79% of the reviewed articles were based on data-driven methods, which rely on analysing and extracting insight from the data, while 21% focused on knowledge-driven methods, which leverage domain expertise to make decisions and draw conclusions. Similarly, Chen et al. [19] found that approximately 70% of the research is on data-driven approaches, with knowledge-based approaches and hybrid data-driven and knowledge-based approaches making up approximately 24% and 6% of the research, respectively. However, in the commercial setting, Granderson et al. [14] evaluated fourteen tools for automated fault detection and diagnostics (AFDD), of which twelve use rule-based algorithms and only three use black-box models to detect faults. Therefore, academia and commercial tools for AHU maintenance appear to be disconnected, indicating that the data-driven approaches found in the literature are often impractical or not reproducible. Thus, a ***practice-theory gap*** emerges that must be addressed with a solution that finds a trade-off between academic novelty and industrial applicability [20]. The inevitable unforeseen challenges of real-world deployment may be categorised as a technical barrier that will need a solution with a high degree of flexibility to overcome.

Considering the motivation for this research and the gaps identified in the published literature, the main research question is;

Main Research Question:

> *"How can reproducible and usable data-driven approaches be developed, implemented and deployed to reduce energy consumption in the industrial setting through the transition to proactive maintenance of AHUs, considering gaps in knowledge, data, operation and practice-theory?"*

This main research question is further developed into four key research sub-questions. The aim of the sub-questions is to address each gap in literature with a published research output to contribute to answering the main research question. They are as follows;

Research sub-questions & associated output:

RQ 1. What distinguishes engineering and data science in terms of problem-solving abilities, and how can the strengths of both disciplines be integrated to develop a proactive maintenance solution for AHUs in the industrial setting?

> **Research Undertaken:** A framework has been developed that combines data science process models (specifically CRISP-DM and DST-map) with engineering-based methods (specifically fault detection and diagnosis and commissioning). The framework is called the Industrial Data Analysis Improvement Cycle (IDAIC) and it guides practitioners through the process of transitioning from reactive to proactive maintenance approaches on AHUs in the industrial setting with the goal of reducing energy consumption. The transition is achieved by firstly leveraging knowledge-based FDD to detect faults that are easily identifiable. This in turn helps to establish a baseline operation that is needed for data-driven FDD techniques which are capable of identifying subtle faults in a proactive maintenance approach.

Research Output: Journal Article 1 – "*Development of a Framework to Aid the Transition from Reactive to Proactive Maintenance Approaches to Enable Energy Reduction*", doi:10.3390/app12136704

RQ 2. How can raw AHU data from the industrial setting be systematically curated into a dataset that enables the development of practical and reproducible data-driven analysis?

Research Undertaken: The raw operational data of an AHU in an industrial partner site is systematically prepared from two heterogeneous sources and contextualised using the Project Haystack naming convention. The dataset addresses the lack of AHU data from real industrial facilities in the literature and encourages the development of hybrid approaches to leverage the five years of unlabelled data.

Research Output: Journal Article 2 – *"A dataset for fault detection and diagnosis of an air handling unit from a real industrial facility"*, doi: 10.17632/8x62ntvrg7.2

RQ 3. How can the operation of a real-world AHU in an industrial facility be assessed to identify the sensor, component and control faults that must be addressed prior to the application of data-driven approaches for the purpose of proactively maintaining the AHU?

Research Undertaken: The developed IDAIC framework has been implemented on an AHU in an industrial partner site to identify existing faults that must be rectified before applying a data-driven approach. This led to a novel extension of the AHU Performance Assessment Rules (APAR) to account for the novel mode of operation caused by the COVID-19 pandemic. Many sensor, component and control faults were identified. For example, a faulty outside air temperature sensor was identified which would have triggered many APAR rules and lead to costly and unnecessary maintenance activity taking place. Furthermore, the detection of an outside air damper stuck in the fully open position from December 2020 to December 2021 caused a significant increase in the mechanical heating demand of the AHU. Considering that there was no mechanical heating demand in the following period of December 2021 to December 2022, resolving this fault led to estimated annual savings of €60,000.

Research Output: Journal Article 3 – *"Implementation of the IDAIC framework on an air handling unit to transition to proactive maintenance"*, doi: 10.1016/j.enbuild.2023.112872

RQ 4. What practical considerations should be taken into account to enable reproducibility in the development of a tool for the purposes of proactively maintaining an AHU in the industrial setting?

Research Undertaken: In collaboration with partners on the DENiM EU project, the IDAIC framework was deployed as a case study in an industrial partner site. This enabled the development of a tool for the continuous monitoring of an AHU, considering practical considerations such as data acquisition, resource limitations, cyber security, and latency issues.

Research Output: Journal Article 4 – *"Deployment of a digital shadow-type tool for the proactive maintenance of an air handling unit in the industrial setting",* In progress

1.3 Scope of the Research

Given the broad scope of research required to integrate developments in engineering and data science, this research focuses on HVAC systems, specifically AHUs. Additionally, as the solutions developed from this research are intended for application in a regulated industry, one-way analysis is employed, meaning that data can only flow from the AHU and cannot be written back to a controller, for example. As a result, the outcomes of this research must be reviewed and approved by a domain expert before any physical activity occurs. To that end, the findings of this research are presented as a decision support aid.

1.4 Structure of the Thesis

Following on from this introduction, **Chapter 1**, the remaining chapters are outlined as follows. **Chapter 2** presents a review of published literature relevant to the research undertaken in this field. The chapter provides an overview of the data analysis field and maintenance approaches, highlights data-driven approaches as the future direction, and proposes knowledge-based approaches to address the current issues, which leads to hybrid approaches being considered as best in class. The gaps in the literature are elicited throughout this review to provide a clear understanding of the challenges in this research area. **Chapter 3** outlines the research methodology that begins with the development of a conceptual framework and leads to the implementation of an industrially applied tool. **Chapter 4** outlines the development of a framework to aid the transition from reactive to proactive maintenance approaches, which addresses the knowledge gap. An integration-type adaptation of a popular data mining process model is proposed to combine the benefits of "learned knowledge" from data mining and "engineered knowledge" elicited from domain experts, as called for by Lepenioti et al. [21]. **Chapter 5** documents the implementation of this framework in a real industrial facility to address both data and operational gaps. Real-world issues are exposed and mitigated in a structured manner to enable the progression of the analysis. **Chapter 6** summarises the extension of the implementation work to a deployed tool. The collaboration and communication between back- and front-end software and facilities engineers is recorded. **Chapter 7** discusses the main outcomes and findings of this research and offers directions for future work.

1.5 Summary of Contributions & Novelty

In summary, the novel contributions of this PhD research are;

- The development of the industrial data analysis improvement cycle (IDAIC) framework to aid the transition from reactive to proactive maintenance approaches
- The curation of a real-world AHU dataset spanning 5 years, which has been systematically prepared and aligned to the Project Haystack naming convention
- The development of a data assessment decision tree for AHUs to identify broken, bad-quality and background data issues
- The extension of the APAR ruleset for novel modes of operation created because of the COVID-19 pandemic
- The deployment of a novel tool for FDD on industrial AHUs

1.6 Research Outputs & Award

The following represents the primary dissemination of the research contained within this thesis:

Peer Reviewed Journal Articles:

M. Ahern, D. T. J. O. Sullivan, and K. Bruton, "Development of a Framework to Aid the Transition from Reactive to Proactive Maintenance Approaches to Enable Energy Reduction," Appl. Sci., 2022.

M. Ahern, D. T. J. O. Sullivan, and K. Bruton, "A dataset for fault detection and diagnosis of an air handling unit from a real industrial facility," *Data Br.*, 2023.

M. Ahern, D. T. J. O. Sullivan, and K. Bruton, "Implementation of the IDAIC framework on an air handling unit to transition to proactive maintenance," Energy Build., vol. 284, p. 112872, 2023.

M. Ahern, D. T. J. O. Sullivan, and K. Bruton, "Deployment of a digital shadow type tool for the proactive maintenance of an air handling unit in the industrial setting", In progress

Conferences

M. Ahern, D. T. J. O. Sullivan, and K. Bruton, "A Systematic Mapping of the recent trends in monitoring the condition of assets for the migration to predictive maintenance," The 3rd International Conference on Mechanical and Digital Manufacturing (ICMDM 2020), Munich, Germany, during June 27-29, 2020

M. Ahern, D. T. J. O. Sullivan, and K. Bruton, "The Application of Industrial Artificial Intelligence in the Manufacturing Industry for Energy Reduction and Fault Detection," Doctoral Symposium at the 6th European Conference of the Prognostics and Health Management Society 2021, online, July 1, 2021

A. Ranade, **M. Ahern**, A. Mumtahina, and K. Bruton, "Industrial Data Analytics Improvement Cycle and it's Application in Industrial Fault Detection and Diagnosis," Nectar track at the 30th Irish Conference on Artificial Intelligence and Cognitive Science (AICS), Cork, Ireland, during December 8th - 9th, 2022

Industry Guide

Provided a synopsis of transitioning from reactive to proactive maintenance strategies for CIM, a building analytics provider, to develop a guide for an industrial audience titled, "An Operations Manager's Guide to HVAC Maintenance"

https://connect.cim.io/operations-managers-guide-to-hvac-maintenance

Award

Awarded a fully funded visit to the Annual ASHRAE Summer Conference in Toronto in June 2022 as part of the ASHRAE Ireland mentorship programme

2 CHAPTER 2 – Literature Review

2.1 Introduction

This chapter introduces, provides context, and documents a critical review of the area of proactive maintenance, to reduce energy consumption in the industrial setting, specific to AHUs. The goal is to understand the gap between traditional reactive approaches and future proactive approaches, while also identifying the barriers which are impeding the progress to close this gap. The proposed solutions in the literature are also discussed to determine the areas in need of further investigation. This chapter is structured as follows; subsection 2.1 provides context and background to the research area, followed by a brief overview of AHUs in subsection 2.2. Subsection 2.3 compares the strengths and weaknesses of different maintenance approaches and how they relate to one another. Subsection 2.4 introduces the data analysis landscape, followed by a review of data-driven approaches and their requirements in subsection 2.5. Subsection 2.6 then introduces knowledge-based approaches as a means of addressing some of the challenges of data-driven approaches. Subsection 2.7 then reviews the integration of expert knowledge and data-driven techniques which is known as a hybrid approach. Subsection 2.8 then highlights the main barriers to hybrid approaches, with conclusions stated in subsection 2.9. The remainder of this subsection, 2.1, provides a brief background regarding the policies that are driving the need for this research.

2.1.1 Background

In 2015 the Member States of the United Nations adopted a shared set of 17 Sustainable Development Goals (SDGs) to ensure a world of peace and prosperity for people and the planet [2]. The first progress report in 2019 emphasized the need for transformative changes and highlighted the interconnections between the different SDGs and the importance of addressing them together. One of the key policy recommendations from the report is the decarbonisation of energy as it is the single largest contributor to climate change [3].

This led to several policy documents emerging, such as the European Green Deal in Europe [22] and the Climate Action Plan 2023 in Ireland [23], that set more specific goals for meeting the SDGs. The European Green Deal aims to make Europe the world's first climate-neutral continent, achieving net-zero greenhouse gas (GHG) emissions by 2050 [22]. While Ireland is committed to achieving a 51% reduction in GHG emissions from 2021 to 2030, then achieving net-zero emissions no later than 2050 [23]. While much attention has been given to transitioning to renewable energy sources, the most expedient and cost-effective strategy to have an immediate impact and reduce the amount of renewable energy needed is to decrease energy consumption. The International Energy Agency (IEA) states that “enhanced efficiency-related measures have the potential to provide one-third of the emissions reductions needed for net zero” [24]. This has led to many government initiatives to drive incremental year-on-year efficiency improvements. For example, in New York City buildings account for approximately two-thirds of GHG emissions [25]. As part of the plan to make the city carbon neutral by 2050, Local Law 97 was passed to reduce the emissions of the city’s largest buildings by 40% before 2030 and 80% before 2050. While federally in the United States, President Joe Biden signed the Inflation Reduction Act of 2022 into law which provides tax incentives for buildings that achieve a reduction of over 25% of total annual energy consumption as outlined in section 13303 [26]. Similarly in the EU, the Energy Performance of Buildings Directive (EPBD) [27] aims to set common standards and promote energy efficiency measures in EU buildings. While in Australia, offices rated using the National Australian Built Environment Rating System (NABERS) have achieved energy savings averaging 42% and have reduced GHG emissions intensity by 53% [28]. Therefore, energy efficiency improvements are recognised globally as a means of making a significant contribution to achieving climate targets.

As a result, established methods have begun to evolve to become more environmentally-centric. One such example in the industrial sector is the emergence of the Lean and Green concept [29]–[31], an evolution of the waste removal and process variation reduction method, lean six sigma [32]. However, the apprehension in these sectors is that further reduction in terms of energy consumption will hamper growth. Therefore, the decoupling of GDP growth from the overuse of environmental resources has been raised as a call to action [3]. In response, further policy has emerged to simultaneously drive growth and reduce energy consumption. In the European Union (EU), the Twin Transition has been proposed to deliver the objectives of the European Green Deal; the need to protect, conserve and enhance the EU's natural capital, and transition to climate neutrality, while also reducing waste [22]. This outlines a digitalisation-based approach which looks to drive growth through new technology such as Artificial Intelligence (AI). This disruptive technology has led to the development of many policies and initiatives internationally as in Table 2-1. The benefits are not only in terms of energy consumption reduction but also extend to many other areas such as predictive maintenance [33], supply chain management [34], production planning [35] and lifecycle management [36]. These digitalisation-based approaches are part of a larger transition to data-driven and more efficient manufacturing known as Industry 4.0.

Table 2-1 Overview of National and Regional AI Strategies

Time of Announcement	Country or Region	National or Regional Initiative/Programs	Ref.
2017	Canada	"Pan-Canadian Artificial Intelligence Strategy"	[37]
2017	China	"New Generation Artificial Intelligence Development Plan"	[38]
2018	European Union	"Coordinated Plan on Artificial Intelligence"	[39]
2018	France	"AI for Humanity"	[40]
2018	Germany	Artificial Intelligence Strategy of Germany Federal Government	[41]
2019	Denmark	National Strategy for Artificial Intelligence	[42]
2019	Japan	AI Strategy 2019 - updated to AI Strategy 2022	[43]
2020	Republic of Korea	National Strategy for Artificial Intelligence	[44]
2020	Finland	Artificial Intelligence 4.0 programme	[45]
2020	United States	"National Artificial Intelligence Initiative"	[46]
2021	Ireland	"AI – HERE FOR GOOD" A National Artificial Intelligence Strategy for Ireland	[47]
2021	United Kingdom	"National AI Strategy"	[48]
2022	Italy	“Strategic Programme on Artificial Intelligence 2022-2024”	[49]

The ultimate strategy is the intersection of three areas; green transition, competitive sectors, and digital transition as in Figure 2-1. This is known as Industry 5.0 in Europe

[50], the sustainability focused evolution of Industry 4.0 as in Figure 2-2. Industry 5.0 is built upon three main objectives; to be sustainable, to be human-centric and to be resilient [51]. In this way, digitalisation-driven solutions should lead to energy efficiency improvements by involving people and allowing for flexibility. This approach is key to contributing to 3 of the European Commission's 6 priorities; "An economy that works for people", "European Green Deal" and "Europe fit for the digital age" [52]. Research conducted by Clancy et al [53] indicates that this type of approach is of value to the manufacturing sector, considering the benefits for effective root-cause analysis of scrap in a foundry process.

Figure 2-1 The drivers transforming European industry [4]

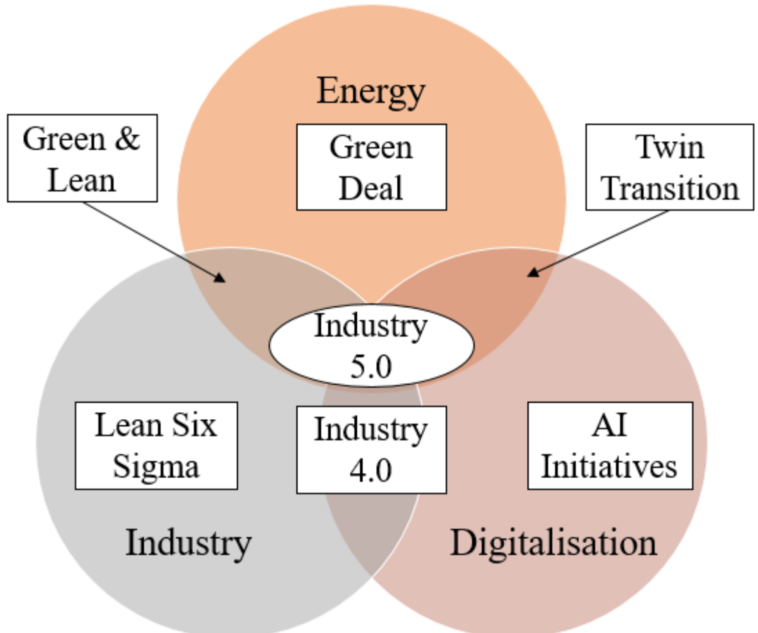


Figure 2-2 Depiction of the policy and strategy landscape for sustainable growth

Similarly in the US, the Energy Information Administration (EIA) projects that the industrial sector's energy use will grow nearly twice as fast as any other end-use sector between 2021 and 2050 [54]. Therefore, the industrial sector faces a significant challenge to attain ambitious sustainability targets, such as doubling the global rate of energy efficiency improvement [2]. In a competitive global market, ensuring sustainability targets are met, while simultaneously driving growth, is a key significant challenge. Incorporating digitalisation approaches such as Information and Communications Technology can enable a 20% reduction in global CO_2 emissions by 2030, which could provide the necessary environmental protection while avoiding a trade-off with economic prosperity [55]. The International Renewable Energy Agency (IRENA) estimates that if manufacturing companies adopted the best available technologies, then energy consumption globally would reduce by about a quarter [56].

One of the key areas that energy reduction may be achieved through digitalisation is the area of proactive maintenance. This area has the potential for high impact if the equipment being maintained is a significant energy user where potentially unnoticed faulty operation leads to decreased efficiency. While the industrial sector is comprised of a variety of subsectors with different significant energy users, indoor facilities have a common, energy-intensive requirement within the buildings sector to provide ventilated working environments. Heating, Ventilation and Air Conditioning (HVAC) systems, which provide fresh air as well as heating and cooling needs, consume an estimated 50% of building energy consumption on average [57], or 14% of energy use across the world [8]. This is an important area to focus on as a UK survey of hi-tech facility managers reports that 34% say HVAC is their site's biggest energy cost [58]. There is a high potential for energy savings in HVAC systems as their self-correcting nature enables compensation of faulty operation, which facilitates the occurrence of unnoticed faults that may account for up to 20% of the energy consumed by these systems [9]. Therefore, one-fifth of the 14% of world energy use consumed by HVAC systems, or 2.8% of global energy consumption, could be reduced by proactively maintaining these systems. HVAC systems therefore need to be well maintained, but maintenance costs, time for repair, and replacement of components that have not reached the end of their useful life must be balanced with the potential energy savings. As noted by O'Donovan et al. [59], "equipment maintenance can exceed 30% of total operating costs, or between 60 and 75% of equipment lifecycle cost [60]". However, maintaining buildings, or building commissioning, is reported to have a significant impact on the overall energy consumption of the building. This area of research has been extensively documented by the Energy Systems Laboratory in Texas A&M. This group coined the term Continuous Commissioning ®, a common sense approach to maintaining building mechanical and control equipment through ongoing monitoring [61]. A study of buildings on the university campus reported average savings of 28% on chilled water, 54% on heating and between 2 to 20% on electrical energy in non-retrofitted buildings [61]. This led to an expected payback in the order of 1 to 2 years [61]. Therefore, selecting an appropriate maintenance strategy is critical to having a cost-effective operation and also contributes to attaining sustainability targets.

2.1.2 Research Objectives

Before implementing proactive maintenance strategies, it is crucial to identify and understand the existing gaps that prevent the proposed strategies from being achieved. The four research questions outlined in subsection 1.2 are explored in detail in the following subsections from a state-of-the-art perspective. Related research aiming to address these gaps are also reviewed to define the specific research areas that this PhD research contributes towards. The objective of the remainder of this chapter is to assess the barriers to proactive maintenance approaches for AHUs in the industrial setting to identify the main steps to be taken to address each barrier.

2.2 Brief Overview of Air Handling Units

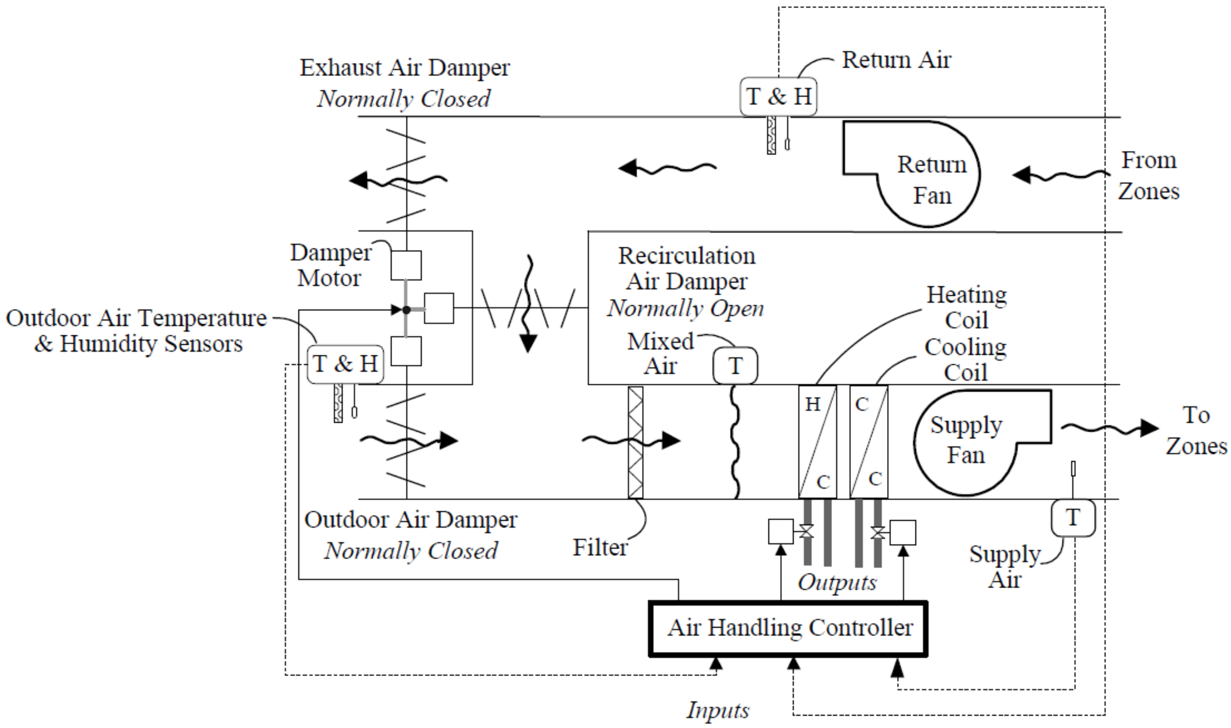


Figure 2-3 Schematic diagram of a single duct air handling unit [62]

Air handling units (AHUs) are a critical component of HVAC systems. Their purpose is to distribute conditioned air throughout a building, or a zone within a building, to provide good thermal comfort and indoor air quality. Many different types exist but the typical single duct AHU shown in Figure 2-3 modulates damper positions to control the fraction of outdoor air that enters the building, while also modulating hot water and chilled water valve positions to control the heating and cooling of the air respectively. This control is provided by a computer-based control system called a building management system (BMS) [63]. The advances in BMS control systems have led to the development of model predictive control (MPC), an approach capable of anticipating a building's response to control requests to then perform the necessary operation efficiently [64]. Therefore, it is a means of enabling better control through energy consumption prediction. In a review of building management systems by Mariano-Hernández et al. [64] there are three types of methods used for building energy use forecasts as in Figure 2-4. Firstly, “white-box” approaches refer to physics-based models that are highly interpretable. The energy consumption of the building is obtained through a transparent process of solving physics equations [65]. Secondly, “black-box” models refer to data-driven models. The energy consumption forecast is obtained through time-series statistical analyses and other advanced algorithms that cannot be interpreted [65]. The third category describes the combination of the two previous models. These hybrid models are called “grey-box” as they combine both white-box and black-box approaches [65].

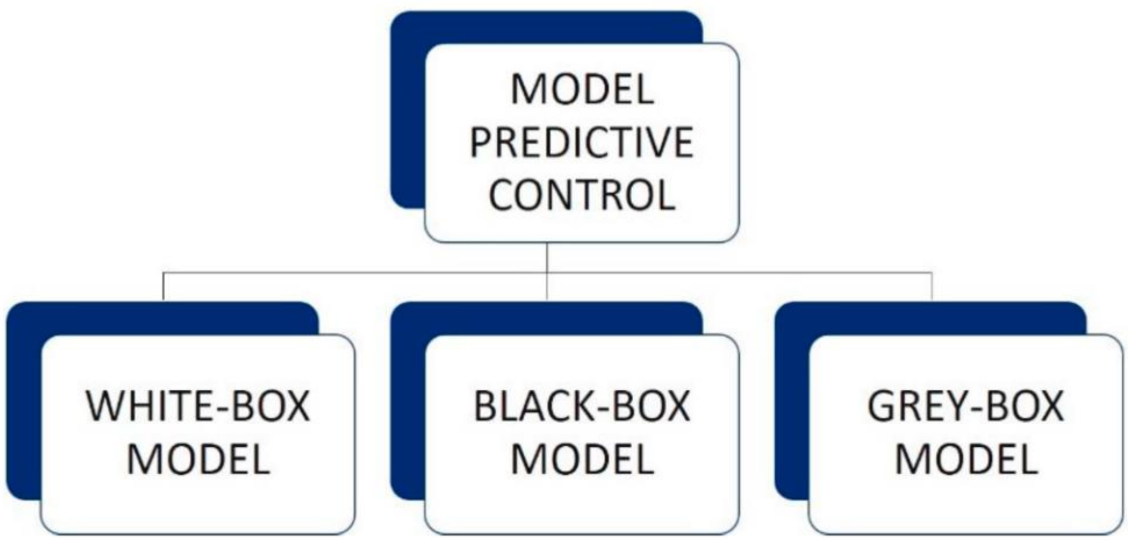


Figure 2-4 MPC-based approaches on building energy consumption prediction [64]

While MPC is an important aspect of managing energy consumption through BMS control, it is only one of four management strategies identified by Mariano-Hernández et al. [64]. The actions of the user side (Demand Side Management) and maximising the use of energy (Optimisation) are also key strategies that contribute to policy targets. However, the final strategy of fault detection and diagnosis is a key means of detecting flaws in both the physical AHU and the BMS control which have a significant impact on energy consumption as discussed in subsection 2.1. In ASHRAE guideline 36, it is suggested that FDD should be incorporated into the control strategy [66]. However, given the strong need for follow-up commissioning [16], a solution that provides oversight of the BMS is still necessary. This in part has led to the emergence of commercial tools such as OpenBlue from Johnson Controls [67] and the PEAK platform from CIM [68] that provide this oversight as a service. In subsection 2.3 an overview of the different types of maintenance approaches is discussed based on the time of fault occurrence.

2.3 Review of Maintenance Approaches

The policy directs future efforts towards sustainable, human-centric and resilient solutions. In the area of maintaining HVAC equipment, this is considered a transition from reactive to proactive approaches. Traditional maintenance practices are reactive in nature with a “fail-and-fix” philosophy [12]. Reactive approaches incorporate the traditional practice of over-maintaining critical equipment in preventive strategies and under-maintaining low-priority equipment in corrective strategies. Proactive approaches differ as they have a “predict-and-prevent” philosophy [12]. They aim to find the optimal time for maintenance by detecting degradation in condition-based strategies, estimating the time until failure in predictive strategies, and guiding maintenance actions in prescriptive strategies. This is a fundamental change in perspective for the area of maintenance.

Traditionally, maintenance is centred on understanding the operation and failure modes of the equipment. However, with the adoption of a digital approach, maintenance will now centre on understanding the patterns in machine-generated data. Therefore, the traditional approach is based on the fundamentals of the engineering domain, while the digitalisation approach is based on the fundamentals of the data science domain. The difference in mind-set between these domains may be contextualised using the potential failure functional failure (P-F) curve in Figure 2-5.

Industrial mechanical equipment degrades over time and maintenance is the process of preserving its condition. A typical degradation pattern is shown in Figure 2-5. After correct maintenance at $t = 0$, or commissioning in the case of HVAC equipment, the health index is at a maximum of 1. Over time, the health index begins to slowly decrease due to wear and tear. The point at which this degradation begins to become a concern is at the potential failure point, t =3, where the equipment may cease to meet its desired function. This may be considered a fault as Isermann states that a fault is “an unpermitted deviation of at least one characteristic property of the system from the acceptable, usual standard condition” [69]. While the fault may not inhibit the overall desired function, it may cause this function to be achieved less efficiently. The condition is likely to deteriorate further after this point until a functional failure occurs at t=4, whereby the desired function can no longer be achieved. Beyond this point, significant issues occur such as unplanned downtime, further loss of efficiency and, depending on the type of failure, safety concerns. However, understanding the degradation pattern, identifying points of fault and failure, and determining the reason for their occurrence can be challenging. To address these issues, Lee et al. [70] identified four types of learning systems in industry: 1) expert’s experience, 2) expert systems, 3) AI/ML and 4) integrated learning, as shown in Figure 2-6. Each of these learning systems forms the foundation of a maintenance strategy which is discussed in the following subsections.

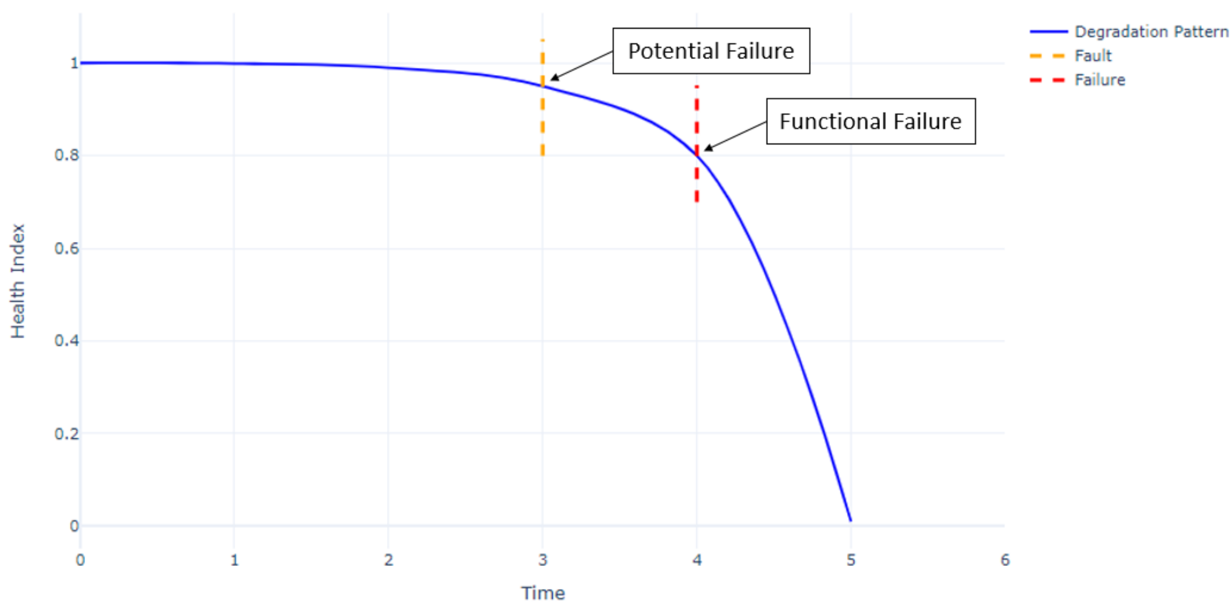


Figure 2-5 Potential failure functional failure curve (P-F curve)

Table 2-2 provides an overview of the origin of each strategy along with the associated type of commissioning activity as it is the preferred term for maintenance in the HVAC domain. As there are both strengths and shortcomings to each type, each of the maintenance strategies are discussed in further detail while also being positioned on the P-F curve as in Figure 2-7.

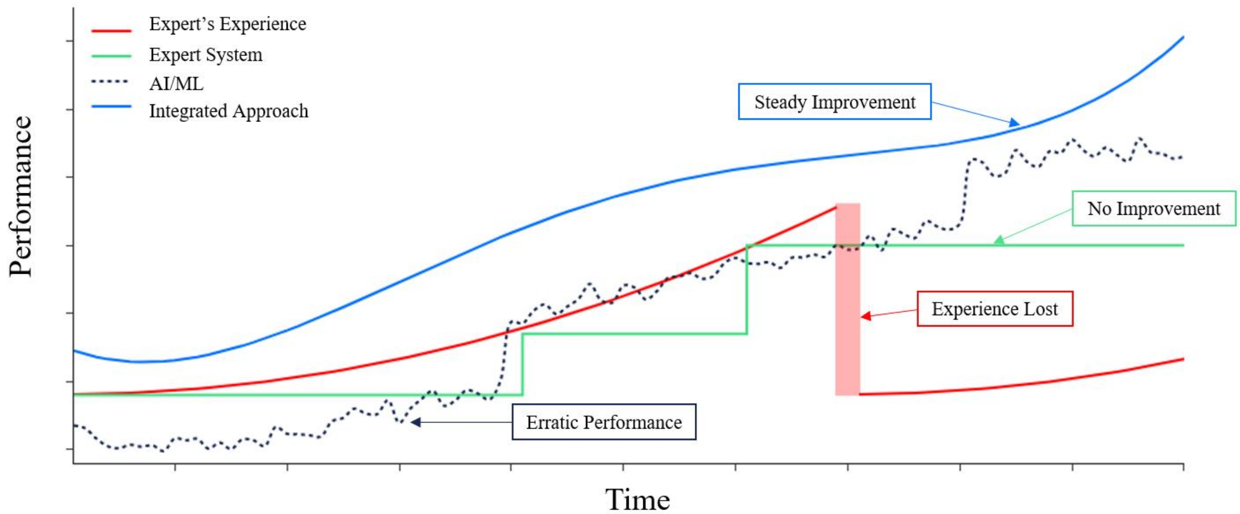


Figure 2-6 Comparison of learning systems (Adapted from Lee et al. [70]*)*

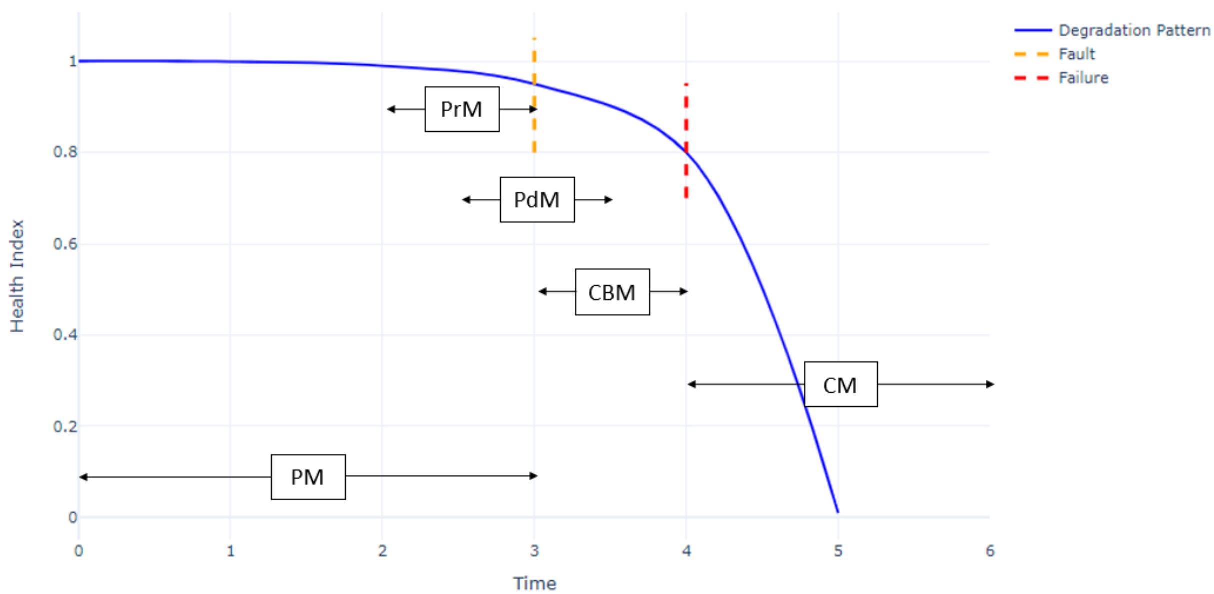


Figure 2-7 P-F curve and maintenance strategy location

Table 2-2 Overview of Maintenance Approaches

Primary Domain	Engineering		Data Science		
Maintenance Approach	Reactive		Proactive		
Maintenance Strategy	Corrective	Preventive	Condition-based	Predictive	Prescriptive
Commissioning	Initial/Retro Cx	RCx	OCx/ CCx/ MBCx		

2.3.1 Corrective Maintenance

Corrective maintenance (CM) aims at restoring equipment to its desired operating condition after a breakdown or failure [71]. As this strategy schedules maintenance after a failure has occurred it is reactive in nature. As broken equipment is being fixed, this strategy is similar to initial commissioning and retro-commissioning which are both first-time processes carried out on a newly constructed building, or an existing facility that was not previously commissioned, respectively [72], [73]. The CM maintenance activity then takes place after t = 4 in Figure 2-7. This has the benefit of maximising the life of the equipment or component which reduces maintenance costs [74]. However, when failure does occur it is usually unplanned [60], resulting in costly downtime and repair [74]. Therefore, the maintenance activity has occurred too late.

This strategy aims at answering the question of “What happened?”, therefore the analysis type is descriptive. The learning system associated with this analysis is expert experience. In this learning system, a domain expert relies on their interpretation of past events to determine “What happened?”. The more events that the expert has been exposed to, the greater their problem-solving ability. However, as in Figure 2-6, the learning process is very slow as it follows an exponential type curve. The knowledge gathered in this type of learning system is difficult to transfer to another person. Therefore when an expert changes positions in an organisation or retires, a significant amount of knowledge is lost as twenty years of experience cannot be transferred to another person in two weeks. Hence, it is not a sustainable approach.

2.3.2 Preventive Maintenance

To avoid the occurrence of a catastrophic failure, a preventive maintenance (PM) strategy may be applied. PM is carried out to decrease the failure probability [71]. This may occur through a scheduled activity which may be either usage-based or time-based [75]. As the equipment may not need maintenance, it is similar to re-commissioning (RCx), a process undertaken on a new or existing building to verify or improve performance [72]. The advantages of PM include increasing equipment availability by focusing on key components [74], performing as convenient, reduction in parts inventory, improving safety, and standardized procedures [76]. While the disadvantages include exposing equipment to possible damage, using a greater number of parts, demanding more frequent access to equipment [76], and increased costs as maintenance may be done prematurely [74]. Therefore, maintenance is performed too early.

As the intervals are scheduled based on manufacturer guidelines or user experience [77], it is based on the same learning system as CM - expert experience. However, the application of the learning system differs as it is based on the question “Why did it happen?” so that measures may be scheduled to prevent reoccurrence. This is a diagnostic type analysis. While PM may be categorised as a proactive approach [78] as it aims at scheduling activity before t=4 in Figure 2-7, its learning system is not sufficiently advanced which leads to maintenance potentially taking place too early which is not optimal in terms of its cost-effectiveness.

2.3.3 Condition-based Maintenance

While typically categorised as a form of PM, condition-based maintenance (CBM) may be considered its own strategy as CBM involves monitoring the condition of the asset to carry out maintenance when necessary rather than following a fixed schedule [77]. This as-needs approach is similar to ongoing commissioning (OCx) or continuous commissioning® (CCx) whereby efforts are undertaken to continually maintain, improve and optimise performance [72]. Specifically, it is known as monitoring-based commissioning (MBCx), ongoing commissioning by way of data monitoring and analysis [79]. Therefore, CBM is a proactive approach as measurements are monitored to determine if a fault is about to occur or already has occurred [59].

To understand the conditions that must be monitored in a CBM strategy, the tacit knowledge of experts needs to be elicited. This may be achieved through an expert system as in Figure 2-7. In this system, the knowledge of experts is captured in the form of rules that compare a variety of measured information. If a rule then exceeds a pre-defined threshold, a fault is determined to have occurred. The advantage of this learning system is that the captured knowledge offers sustained performance over time. However, this learning system is limited to the knowledge of the expert and may need to be updated regularly to reflect changes in dynamic industrial environments. Therefore, it remains a diagnostic form of analysis. Using the CBM strategy, maintenance activity is performed after the fault ($t > 3$) and before failure ($t < 4$) in Figure 2-7. This may be considered the optimal region for maintenance activity to take place.

2.3.4 Predictive Maintenance

The next advancement is the predictive maintenance strategy (PdM). As defined by Basri et al. [78], "In PdM, repetitive or high-risk failures are studied using historical data detailing occurrences of a machine's operational failures and then maintenance is conducted during its operation, based on the condition of the monitored component". This is an evolution of CBM as the monitored condition is used to estimate the remaining system life [77]. This is also similar to ongoing commissioning. The additional benefit of this approach is that scheduling activities are improved through the estimation of the remaining useful life of the system. Benefits identified by Achouch et al. [11] include minimising the occurrence of unscheduled downtime, maximising asset uptime, ensuring minimal disruptions to productivity as predictive maintenance activities can be carried out on running assets, optimising the time spent on maintenance work, optimising the use of spare parts, and improving asset reliability. However, given that the models needed to predict maintenance requirements can be complex, it may be time-consuming to implement a predictive maintenance strategy [74].

PdM is built on the ability to learn from historical data. The analysis is predictive, which is a more advanced form than descriptive or diagnostic analysis, focusing on the question of "What will happen?". Therefore, an evolution in the learning system needs to occur. The AI techniques that are driving the digitalisation transformation provide this evolution, specifically the machine learning (ML) subset. The typical performance of these techniques over time is graphed in Figure 2-6. In the beginning, these techniques generally perform poorly as there are not many examples available that AI/ML can learn from. The performance then begins to improve as more events occur from which the AI/ML can learn. While the advantage of this learning system is that it can extend beyond the knowledge of experts, its performance can be erratic. This instability may be caused by the dynamic conditions experienced in the industrial setting such as changeable working conditions and diversified data [80]. The maintenance activity remains in the optimal region with the added benefit of improved scheduling.

2.3.5 Prescriptive Maintenance

The final and most advanced strategy is a Prescriptive Maintenance strategy (PrM). This strategy aims to answer the question "How can it be made happen?". It does this by "providing actionable recommendations for decision-making and improving and/or optimising forth-coming maintenance processes" [81]. It may also be considered a form of ongoing commissioning. This is the strategy with the highest degree of maturity which involves the use of complex methods [81]. This approach therefore requires the most advanced learning system which is the integrated, human and AI approach. In this approach, the domain expert works with AI/ML techniques to seek to find the best course of action for the future [21].

2.3.6 Comparison

The benefits and shortcomings of each strategy are summarised in Table 2-3. The main differences between reactive and proactive approaches are as follows. Firstly, reactive maintenance strategies are primarily concerned with the past, "what happened" and "why did it happen". While proactive approaches are primarily concerned with the future, "what will happen" and "how can we make it happen". CBM, while categorised as a proactive approach in Table 2-2, is primarily concerned with the present moment which suggests it may offer a bridge between both mind-sets. Secondly, the information requirements of the reactive strategies are much less than the proactive strategies. The reactive strategies utilise engineering knowledge that may be gathered through experience while the proactive strategies have a high dependence on data from which to learn. Thirdly, the learning systems in proactive approaches are more sustainable than reactive approaches. Tacit knowledge gained by domain experts is easily lost and difficult to transfer, while proactive approaches are capable of retaining learned knowledge. This suggests that proactive maintenance approaches are more suited to fulfil the goals of EU policy to be sustainable, human-centric and resilient.

Table 2-3 Comparison of the benefits and shortcomings of each maintenance strategy

Type	Activity Time	Learning System	Benefits	Shortcomings
CM	Too late	Expert Experience	• Low upfront costs • Addresses issues only when they occur • Appropriate for low-value or non-critical equipment	• Can lead to unexpected downtime • Can result in more frequent repairs and higher costs in the long run • Can shorten equipment lifespan
PM	Too early	Expert Experience	• Reduces unexpected downtime and equipment failures • Increases equipment lifespan • Enables planned maintenance and repairs • Decreases maintenance costs in the long run	• Can result in unnecessary maintenance and downtime • May not solve all potential issues • Can be expensive upfront
CBM	Optimal	Expert System	• Maintenance performed only when needed • Reduces unnecessary maintenance • Increases equipment lifespan • Greater problem detection capability	• Requires specialized equipment and training • Not appropriate for all equipment • May not detect all potential issues
PdM	Optimal	AI/ML	• Can detect potential issues before they result in equipment failure • Reduces unexpected downtime and equipment failures • Increases equipment lifespan • Enables planned maintenance and repairs	• Requires specialized equipment and training • Can be expensive upfront • May not detect all potential issues
PrM	Optimal	Integrated Learning	• Provides specific recommendations for maintenance and repairs • Can optimize maintenance schedules and procedures • Reduces unnecessary maintenance and downtime	• Requires advanced data analysis and modelling • Can be expensive to implement and maintain • May not be appropriate for all equipment

2.3.7 Conclusion

In this subsection, 2.3, the issues of understanding the degradation pattern, determining the points of fault and failure, and determining the cause of their occurrence were introduced. The learning strategies used by each approach were then discussed, such as using expert systems to identify the fault point in CBM and using AI/ML to identify the degradation pattern to determine the time until failure in PdM. However, monitoring the degradation pattern is inherently difficult as it may take many different forms. For example, a degradation pattern for each of the three other similar assets to that in Figure 2-5 is shown in Figure 2-8. Each pattern begins at a different health index which could be caused by improper commissioning or differences between the assets such as age. Assuming that the fault and failure points are the same for each pattern which remains at t = 3 and t = 4 respectively, it is difficult to determine the fault point for patterns B, C and D from degradation pattern A. Further complexity is added when other differences are considered such as different operating conditions, different equipment types and different fault types. The understanding of this operational detail of mechanical equipment and understanding of the AI/ML techniques is not typically possessed by the same person in an industrial facility. This is a ***knowledge gap***. Therefore, to implement a proactive maintenance strategy, the understanding of the domain area must be integrated with data-driven techniques to address the questions of “At what point does a fault occur?” and “What is causing the fault to occur”?

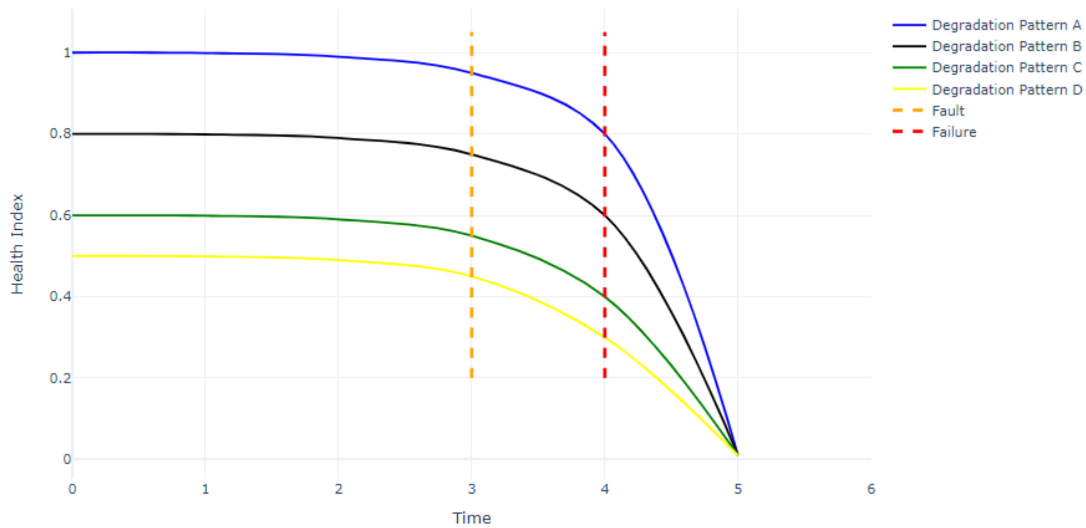


Figure 2-8 Comparison of different potential failure curves

In the area of HVAC maintenance, the identification of the fault point and the reason for its occurrence is a mature field known as Fault Detection and Diagnostics (FDD). FDD has been applied in the area of HVAC and AHUs for over 35 years. Over that time the literature has evolved from primarily focusing on expert-based learning systems using thermodynamic relationships to analyse faults [82], to AI/ML learning systems in more recent times. Many different types of FDD approaches have been developed over this time, with many FDD classification schemas also proposed. The more recent reviews of this research area tend to classify FDD methods as data-driven, knowledge-based (including model-based and physics-based), and hybrid approaches [19][83]. Subsections 2.5 - 2.7 discuss each of these approaches in further detail. However, before reviewing the latest literature on AHU FDD, the main philosophies that guide the implementation of data analysis, of which FDD is a subset, must be understood. Therefore, a brief overview of the seminal work in data science is provided in the following subsection.

2.4 Data Analysis Concepts and Adaptations

This subsection 2.4 provides an overview of the field of data analysis, including its key concepts and adaptations that have emerged over the years as the field has progressed. During this progression, the meaning of some terminology has evolved while other terms are often used interchangeably. For clarification purposes, the relevant Oxford English Dictionary definition of terms frequently used throughout this thesis is provided along with examples of their implementation in Table 2-4. The terms are sequentially ordered based on their definitions. For example, a framework provides the conceptual foundation for analysis that may be carried out following a process model. A method would then refer to the approach taken to perform the analysis task, which is specifically carried out using an algorithm. Finally, the analysis is delivered, or deployed, as a tool.

As the key concepts and adaptations in the data analysis field are the focus of this subsection, the relevant literature at the framework and process model level is discussed. More detailed research at the method and algorithm level is discussed in subsection 2.5, specifically concerning AHU FDD.

Table 2-4 Overview of data analysis terminology as defined by the Oxford English Dictionary [84]

Level	Definition	Example
Framework	An essential or underlying structure; a provisional design, an outline; a conceptual scheme or system	KDD
Process Model	A diagram or chart outlining the various steps involved in a particular process or set of processes	CRISP-DM
Method	A way of doing anything, esp. according to a defined and regular plan	Regression Analysis
Algorithm	A precisely defined set of mathematical or logical operations for the performance of a particular task	Ordinary Least Squares
Tool	Any item of software used as the means of accomplishing a specific task	Commercial Software

In 1996, the Cross-Industry Standard Process for Data Mining (CRISP-DM) was conceived, providing a “blueprint for conducting a data mining project” [85]. Data mining may be described as “the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner” [86]. CRISP-DM provides a structured six-phase approach with highly flexible transitions between the phases, denoted by the multiple arrows and cyclical nature outlined in Figure 2-9. CRISP-DM adequately satisfied the needs of data analysts and became established as the de facto standard for industry [85]. In 2000, Shearer envisaged that extensions and improvements were to be expected [85], and this evolution was analysed in 2019 by Martínez-Plumed et al. [87].

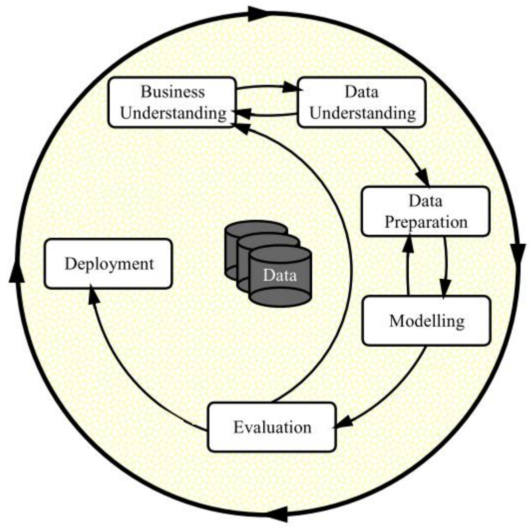


Figure 2-9 The Cross Industry Process for Data Mining (CRISP-DM) process model [88]

The evolutions satisfy specific needs while retaining the principles and ideas of CRISP-DM or Knowledge Discovery in Databases (KDD). In 1996, Fayyad et al. [89] believed that "KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process". However, in more recent times, data mining is often used as a synonym for KDD [87], therefore data mining is the preferred term. Figure 2-10 outlines some of the most influential methodologies in the development of CRISP-DM, along with some adaptations that it has inspired. Figure 2-10 is an adaptation of the evolution illustrated by Mariscal et al. [90] which was released in 2010. Mariscal et al. [90] name KDD and CRISP-DM as the two main approaches that are referred to as the "canonical" methodologies by Martínez-Plumed et al. [87], depicted in grey in Figure 2-10. KDD along with SEMMA (Sample, Explore, Modify, Model and Assess) and Two Crows process models. These process models were developed by the SAS Institute and Two Crows Corporation respectively. Other relevant methods include 6-sigma, a quality management method aimed at eliminating defects through five phases

(Define, Measure, Analyse, Improve and Control), and the 5A's (Assessment, Ask, Acquire, Appropriate, and Apply) developed by the US Army to ask the right questions and obtain the right data before applying data mining techniques. More recent adaptations appear to originate from CRISP-DM, addressing a variety of specific needs such as the project management need in IBM (ASUM-DM [91]), the need for collaboration between geographically diverse groups (RAMSYS [92]) and the need for context change (CASP-DM [93]).

More recently in 2020, Plotnikova et al. [94] performed a systematic literature review of data mining adaptations, categorizing the derivative work as modification, extension and integration-type adaptations. Modifications aim at solving a particular problem in a given case study, while extensions are aimed at altering data mining methodologies to account for specialized environments and incorporate context awareness [94]. In contrast, integration-type adaptations are aimed at combining data mining methods with other domain frameworks, methodologies and concepts [94]. Before 2008, most data mining methods (CRISP-DM) were applied "as is", whereas the use of adapted data mining methodologies slowly began to take precedence in recent years [94]. Therefore, a key finding of the Plotnikova et al. [94] review is that data mining methodologies now need to be framed as part of "a broader ecosystem of methodologies", rather than the traditional, isolated implementation [94].

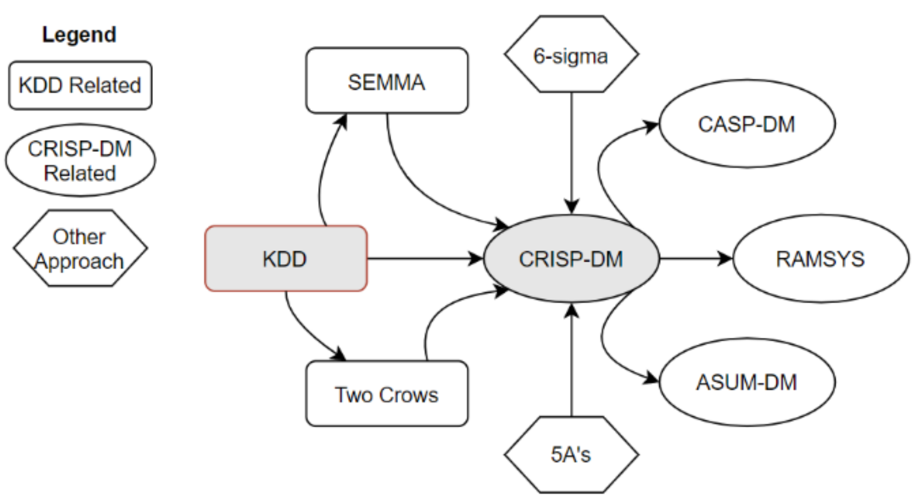


Figure 2-10 Evolution of Data Mining and Data Science models and methodologies based on the study of Martínez-Plumed et al. [87].

A change in the data mining landscape is also observed by Martínez-Plumed et al. [87], noting that the term "data science" is now preferred to the use of the term data mining. Martínez-Plumed et al. [87] make the argument that "the key difference (we perceive) between Data Mining twenty years ago and data science today is that the former is goal-oriented and concentrates on the process, while the latter is data-oriented and exploratory". The term exploratory data analysis (EDA), a closely related term to data mining, was first introduced in 1977 by John Tukey [95] to describe the detective-like analysis of data—that is to say, to not take the data at face value or as absolute truth. Therefore, it is important to understand the data and the problem that needs to be solved to determine if the analysis is goal-oriented or data-oriented.

The field has seen a general progression from goal-based data mining to exploratory-based data science. However, the two are very much interlinked as in Table 2-5. Data mining is generally performed using the CRISP-DM process model while exploratory-based activity is less defined, however, some solutions have been proposed such as the Data Science Trajectories (DST) map in Figure 2-11. Both techniques are useful in the context of proactive maintenance in the industrial setting but the specific purpose of each approach must be understood to know how and when it should be used.

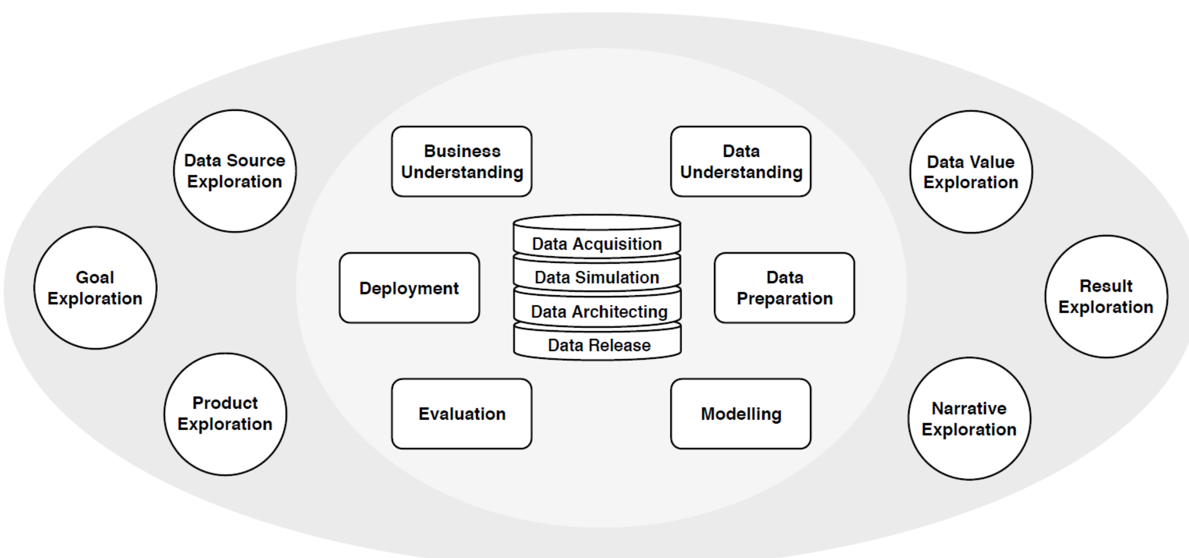


Figure 2-11 The Data Science Trajectories Map (DST Map) proposed by Martínez-Plumed et al [87]

Table 2-5 Comparison of data science and data mining process models in the literature adapted from Azevedo et al. [96]

Process Model	**KDD**	**SEMMA**	**CRISP-DM**	**DST Map**
Activity	***Goal-directed***	***Goal-directed***	***Goal-directed***	***Exploratory***
Understanding the application domain	Pre KDD	---	Business Understanding	Goal Exploration
Acquiring the relevant data	Selection	Sample	Data Understanding	Data Source Exploration
Searching for unanticipated trends and anomalies	Pre-processing	Explore	Data Understanding	Data Value Exploration
Converting raw data to a form ready for analysis	Transformation	Modify	Data Preparation	---
Assessing the health state of an asset	---	---	---	---
Selection and application of an analysis technique	Data Mining	Model	Modelling	Result Exploration
Relating the results back to the goals	Interpretation/Evaluation	Assessment	Evaluation	Result Exploration
Implementation of the analysis to provide value to the user	Post KDD	---	Deployment	Product Exploration

A simplified overview of the difference between goal-oriented analysis and exploratory-based analysis is represented in Figure 2-12. In goal-oriented analysis, there is a particular fault of interest that is known to be a fault of concern. This analysis then takes the form of analysing the data with the specific goal of detecting this fault. An example of the proactive maintenance approach this leads to is the identification of a leaking cooling coil valve. In exploration-based analysis, there is no particular fault of interest. In this approach, the data is analysed to detect abnormal behaviour that is indicative of the occurrence of a hidden fault. An example may be the detection of an abnormal temperature sensor reading at a particular set of control signals that may lead to the identification of a hidden fault that has been masked through compensating control. In both of these approaches, domain knowledge is required. The expert must communicate the fault of interest in the goal-oriented approach given the lack of labelled datasets, and the expert must interpret the unexpected deviations in the exploratory-based approach to diagnose the cause of the anomaly. Both approaches should therefore be implemented, the known failure modes should be analysed using a goal-oriented approach and the hidden faults should be analysed using an exploratory-based approach.

However, CRISP-DM and the DST map were not developed specifically for the data-driven analysis of AHUs in the industrial setting or for proactive maintenance. They are generic solutions for a wide range of potential data analysis scenarios. Therefore, appropriate adaptations need to be made, which is common as discussed by Plotnikova et al. [94]. As there is a need to adapt CRISP-DM and DST map to include additional information, an integration-type adaptation is required. These process models are often adapted to fulfil the requirements of the user. The collaboration of traditionally different disciplines (engineering and computer science) "is necessary to drive progress" [97], proven by the increasing number of studies analysing the integration of domain-specific frameworks, methodologies, and concepts with data mining methodologies [94]. This is a difficult task, however, as the historical lack of emphasis on data analytics has created "a skills gap when it comes to a data centered-mindset amongst the manufacturing workers" [98].

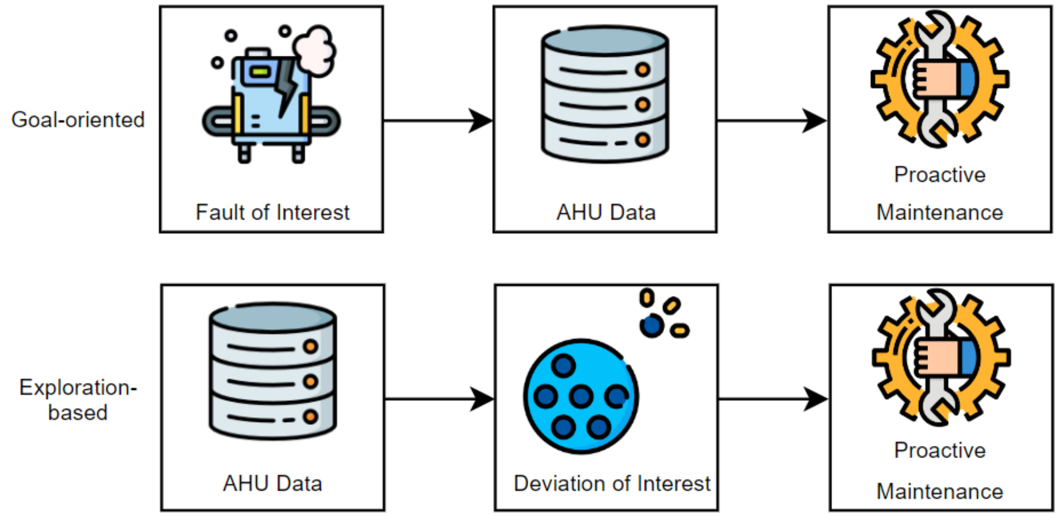


Figure 2-12 The two main data analysis approaches to proactive maintenance

Proactive maintenance in the industrial setting has been researched extensively by the IMS center [99]. Professor Jay Lee, founding director of the IMS center, coined the term “Industrial AI” to describe the application of artificial intelligence in the industrial domain [80]. Concerning goal-based and exploration-based activities, they have proposed that problems be classified as visible (those known to domain experts) and invisible (those unknown to domain experts). In this way, visible problems may be addressed using a CRISP-DM-based approach and invisible problems may be addressed using a DST map based approach. This provides a means of creating a singular approach that incorporates the value of domain experts and data-driven methods. Subsection 2.5 further reviews the application of data analysis in the industrial domain.

2.5 Data-Driven Approaches

The EU proposed move to digitalisation to attain climate targets is a transition to using data-driven approaches to drive sustainability objectives. These approaches have the capability to learn from the vast quantity of collected data to uncover hidden trends and clusters that are indicative of the occurrence of a fault [17], which contributes to energy inefficiency in AHUs. In the AHU maintenance context, learning previous degradation curves, such as those in Figure 2-8, may be used to detect the early signs of repeat occurrences.

The 2021–2027 Horizon Europe funding programme prioritises six key enabling technologies: advanced manufacturing, advanced materials, life-science technologies, micro/nano-electronics and photonics, artificial intelligence, and security and connectivity [100]. Among these technologies, AI is the most relevant for proactive AHU maintenance. AI drives the transition towards more sustainable practices [101], and possesses exceptional learning capabilities, enabling it to extract valuable information from data for maintenance decision-making.

AI is a branch of computer science that aims to mimic the brain's ability to learn from experiences. A subset of AI is machine learning (ML) where the main definition from 1959, but which is still valid today [97], is allowing computers to solve problems without being specifically programmed to do so [102]. AI and ML give machines the ability to learn and decipher the solution to various problems that would normally rely on human intelligence. The main benefit of AI techniques is that it can be far easier to train a system to identify abnormal behaviour by presenting it with examples of desired input-output behaviour, rather than manually modelling or programming for every possible input [103]. ML has been successfully applied to perform tasks such as computer vision, speech recognition, natural language processing and robot control [103] in sectors such as marketing, finance, telecommunications and network analysis [104]. This technology could enable manufacturers to get greater insight into the performance of their assets by clearly identifying trends and characteristics that have not yet been fully recognized [80]. The insight and evidence contained in the data also enable predictive analysis, which is a key benefit of AI [80]. Through the exploration of complex relationships in the data, new knowledge is generated that may be iteratively improved sustainably and continuously to increase the efficiency of industrial systems [80]. There are four basic paradigms in ML as in Figure 2-13; supervised, unsupervised, semi-supervised and reinforcement learning [103].

The upcoming subsections introduce these different ML paradigms and provide examples of their application in AHU FDD found in the literature. Subsection 2.5.4 then discusses the strengths and shortcomings of these methods, followed by a concluding section.

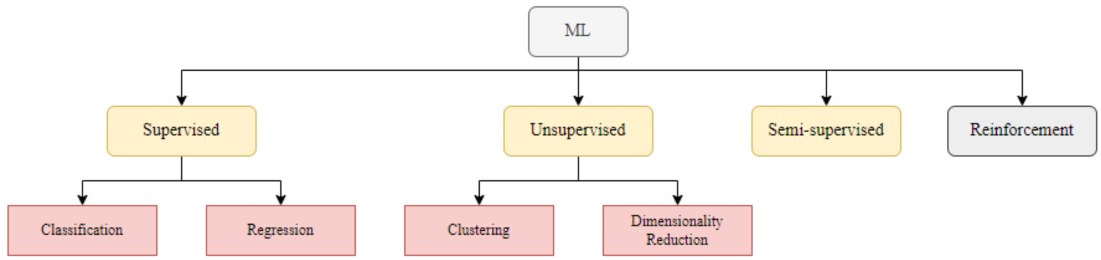


Figure 2-13 Categorisation of ML paradigms

2.5.1 Supervised

Supervised learning, the most widely used method [103], learns a function that maps an input to an output based on labelled training data. The most common supervised learning algorithms include neural networks, support vector machines, decision trees, logistic regression and Naïve Bayes classifiers [20]. The output variables may be either categorical or continuous and are known as classification tasks (discrete data values) and regression tasks (continuous data values), respectively. Examples in the industrial context outlined by Bertolini et al. [20] include process failure detection (classification) and physical property prediction such as thickness or surface roughness from parts processed by a numerical control machine (regression).

In the area of HVAC FDD, it is the most common ML strategy used by researchers and developers, although labelled datasets are not easy to obtain [19]. Supervised approaches are often highly interpretable which creates a sense of reliability [105]. The most commonly used methods in the literature for classification include Support Vector Machine (SVM) and Decision Trees (DT), which classify new data as faulty or fault-free [106]. While the use of regression techniques which predict a continuous variable and compare it to the baseline operation of the system to detect faults include Support Vector Regressions (SVR) and Neural Networks (NN) [106]. Multi-Layer Perceptrons (MLPs) and their specific deep learning (DL) derivatives such as Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have also been found to be popular [105]. An overview of some of the most recent methods applied to AHUs in literature is provided in Table 2-6 which is a replication of the table produced by Matetic et al. [105] along with other notable works.

Table 2-6 Selected research on supervised learning applied to AHUs

Algorithm	Data	Sample	Software	Year	Ref.
HRF-SVM	ASHRAE 1312-RP	1m	Not Specified	2021	[107]
MCNN	Experimental	Not Specified	Not Specified	2021	[108]
2-D CNN	Synthetic	Not Specified	TRNSYS	2021	[109]
MLP	Experimental	1m	MATLAB & TRNSYS	2021	[110]
SAE	ASHRAE 1312-RP	1m	Python & Tensorflow	2021	[111]
RACNN	Experimental	35s	Not Specified	2021	[112]
Bayesian inference	Synthetic	Not Specified	TRNSYS	2021	[113]
DTO-DRNN	Lawrence Berkeley National Laboratory Dataset	1m	Python	2021	[114]
CC-RF	Experimental	35s	Not Specified	2021	[115]
DBN	Historical BMS data from a school	1h	MATLAB	2020	[116]
ARX, RF	Simulated	Not Specified	IDA ICE	2020	[117]
Ensemble (WGAN-SVM)	ASHRAE 1312-RP	1m	Not Specified	2019	[118]
HMM	ASHRAE 1312-RP	1m	Not Specified	2019	[119]

2.5.2 Unsupervised

Unsupervised learning, on the other hand, does not receive any output information. Its goal is to build representations of the input that can be used for decision-making, predicting future inputs, efficiently communicating the inputs to another machine, etc. [120]. Unsupervised approaches do not attempt to predict output, but instead seek to find hidden correlations in the dataset [106]. Ghahramani [120] suggests that unsupervised learning can be thought of as finding patterns in data that would otherwise be considered unstructured noise. Classic examples include clustering and dimensionality reduction. For FDD, these methods are easier to develop and deploy as only fault-free data is required [106]. Zhao et al. [17] classified these approaches as association rule mining-based, clustering-based, discord detection-based, and principal component analysis-based. While Matetic et al. [105] found examples of K-means clustering and FP-growth which is an association-based algorithm that provides information about the underlying rules that make up the structure of the data. An overview of the methods identified by Matetic et al. [105] for AHUs is provided in Table 2-7.

Table 2-7 Selected research on unsupervised learning applied to AHUs

Algorithm	Data	Sample	Software	Year	Ref.
Clustering	Not Specified	Not Specified	MATLAB to Python	2021	[121]
Unfold-PCA	Not Specified	15m	Not Specified	2021	[122]
TBKSFA	ASHRAE 1312-RP	1m	Not Specified	2021	[123]
Clustering	Not Specified	1h	Not Specified	2020	[124]
CWGAN, ensemble (SVM-RF-DT)	ASHRAE 1312-RP	1m	Not Specified	2020	[125]
PCA	Synthetic	Not Specified	HVACSIM+	2020	[126]

2.5.3 Semi-supervised and Reinforcement

Lastly, semi-supervised approaches are those methods that contain properties of both supervised and unsupervised methods. Usually, a small portion of labelled data is used to improve the learning ability of unsupervised methods. Matetic et al. [105] note that Generative Adversarial Networks (GAN) is a popular method which can provide more data and balance classes by generating a synthetic dataset. An overview of the methods identified by Matetic et al. [105] for AHUs is provided in Table 2-8.

Table 2-8 Selected research on semi-supervised learning applied to AHUs (blank cells represent "not specified")

Algorithm	Data	Sample	Software	Year	Ref.
LSTM-SVDD	Synthetic	Not Specified	OpenModelica	2022	[127]
self-MLP	ASHRAE 1312-RP	1m	R & Keras	2021	[128]
MLP	ASHRAE 1312-RP	1m	Not Specified	2021	[129]
TARM, CART	Not Specified	Not Specified	Not Specified	2020	[130]
OC-SVM	Not Specified	Not Specified	Not Specified	2020	[131]

Reinforcement learning is another type of ML that differs by interacting with its environment and producing actions. These actions change the state of the environment which causes the algorithm to receive rewards or punishments. According to Ghahramani [120], the goal of the algorithm is to maximize the rewards and minimize the punishments it receives over its lifetime. It is closely related to control theory in engineering. As it involves interacting with the physical environment it is outside the scope of this research.

2.5.4 Comparison

A data-driven method "does not require any intervention of human knowledge or physical models, and it only needs real system operational data" [132]. The unknown nature of how the results are obtained cause these methods to be referred to as black-box. There are two main approaches; supervised and unsupervised. Supervised methods are trained on labelled datasets to predict the class of an unlabelled data point. They are predominantly classified as regression-based or classification-based [17]. They are the most popular approaches in literature given the availability of labelled experimental and simulated datasets. However, it is difficult to obtain labels in real-world industrial datasets. One way of overcoming this issue is to train the data-driven model on simulated data and then apply it to real building data. This approach was tested by Huang et al [133] but the authors found that models trained on simulated data are not generalized to be applied for real building data fault detection, leading to less than satisfactory results. The generalisation of these approaches, and the ability to perform well on previously unseen data, is a challenge often noted in systematic reviews [8], [19], [105], [134]. On the other hand, unsupervised methods do not require labels and discover knowledge within the dataset. However, much of the knowledge discovered is useless and they are not good at discovering complex relationships between multiple variables in comparison to supervised approaches [17]. There is also a third type, semi-supervised approaches, which share properties with both of these approaches. While they aim to address the data labelling issue, they may be sensitive to the noisy data that is to be found in the industrial setting.

The success of each of these approaches appears to be highly influenced by the data as in Table 2-9. While data-driven approaches have great learning capabilities, they seem to lack problem understanding. This may not be an issue in domains with high quantities of high-quality data as the datasets sufficiently describe the problem under investigation. However, the literature suggests that this is not the case in the data-driven AHU FDD area as the data is insufficient and difficult to access [8], the data that is available is of inadequate quality [19], and the research is performed on the same datasets with a theoretical focus that is not reproducible [83].

While ML is one of the key enablers to evolving a traditional manufacturing system up to the Industry 4.0 level [135], ML is not applicable to every industrial problem. Lee et al. [12] recommend that systems with differing complexity and uncertainty levels should have different maintenance strategies, and by extension, do not need ML solutions. In the case of HVAC systems on industrial sites, there is low availability of facilities personnel to tend to these assets, often outnumbered by 20 AHUs to one technician [136]. Therefore, costly external consultancy is required to ascertain the cause of complex issues [72] and prioritize the list of issues that need to be resolved. For this reason, the ideal solution is a proactive, data-driven maintenance strategy. However, while data-driven methods perform well in large-scale and complex systems, they require high-quality and sufficient training data [132], which has not been a priority in the past.

Table 2-9 Comparison of the strengths and shortcomings of data-driven techniques (Adapted from Zhao et al.)

Strengths	Shortcomings
Supervised Methods	
• Domain Knowledge not required • Higher accuracy than knowledge-based • Works with some important variables missing	• Mostly unexplainable • Performance is poor if operation is outside the range of conditions in the training dataset • Quality of data has a great impact on performance • Data labels are needed. Time-consuming
Unsupervised Methods	
• Labels not required - potential for wider application • Discovered knowledge always has physical meaning • Can discover hidden knowledge	• Most of the knowledge discovered is useless. Time-consuming to find valuable knowledge • Depends on the quality of the data significantly • Not good at discovering complex relationships between multiple variables
Semi-supervised Methods	
• Leverages large amounts of unlabelled data • Potentially reduces labelling costs • May improve model accuracy • May improve model robustness	• Requires careful selection of unlabelled data • May require more complex model architectures • May be more sensitive to noisy data • May be less interpretable

2.5.5 Conclusion

Transitioning to data-driven, proactive and energy-efficient practices is hampered by real-world data issues. As discussed in subsection 2.3.1, in reactive approaches engineers learn from experience which often takes the form of unstructured or undocumented information such as meetings from past projects. However, data-driven approaches require all the necessary data for the specific task it aims to achieve. As industry currently adopts reactive approaches with low data requirements, there is little emphasis on data collection and analysis. This is supported by a 2022 survey of hi-tech facility managers which reports that 62% of managers admit to being insufficient in terms of their collection and analysis of HVAC data [58]. Research efforts should therefore be directed towards developing techniques that address this ***data gap***, but this is currently not the case. According to Zhao et al. [18], most studies in the literature utilize experimental or simulated data from laboratory tests such as ASHRAE RP-1312 [137] or those made available by the Lawrence Berkeley National Laboratory [138]. Additionally, Chen et al. [106] found that 48% of HVAC literature utilises laboratory data, 20% utilises simulated data, and only 32% uses real data. While these datasets address a severe lack of publicly available benchmark datasets, they are not representative of real-world datasets. Given the generalisation issue of these approaches [8], [19], [103], [133], [136], the lack of datasets representative of the issues encountered in the industrial setting is a barrier that must be addressed.

The same survey of hi-tech facility managers reports that 40% of facilities are stuck in reactive mode for their HVAC maintenance strategy [58]. This means that reactive maintenance strategies, such as CM and PM, are not facilitating the transition to proactive maintenance approaches in their current form. Therefore, these reactive, knowledge-based approaches must evolve to fulfil the data requirements of proactive maintenance strategies. The following subsection 2.6 discusses the different types of knowledge-based approaches to determine how this evolution may take place.

2.6 Knowledge-based Approaches

As discussed in the review of maintenance approaches, the reactive strategies are based on the expert's experience learning system. This learning system then evolved to capture the tacit knowledge of experts through expert systems to develop a CBM strategy. However, there is a disconnect to the AI/ML learning system that enables the advanced proactive strategies of PdM and PrM. The different types of expert systems or knowledge-based approaches are discussed in this subsection 2.6 to close the gap to AI/ML.

In the past, knowledge-based methods have been defined as the qualitative part of model-based and data-driven methods. In more recent literature, knowledge-based approaches typically relate to rule-based algorithms, while physics-based approaches refer to detailed physical models [105]. As both of these approaches rely on capturing experts' experience or knowledge, they originate from expert system learning systems. Therefore, these approaches may be considered as two types of expert systems; shallow knowledge (rule-based) and deep knowledge (physical modes) [139]. Both are discussed in this subsection in the context of evolving reactive knowledge-based techniques to fill the data gap to aid the transition to data-driven approaches.

2.6.1 Shallow Knowledge Systems

Rule-based systems, or shallow knowledge, are white-box models meaning that they are highly interpretable. They detect faults through the use of "IF-THEN" rules, checking if particular conditions are met. In 2008, Fan et al. [140] noted that most knowledge-based methods are rule-based expert systems, which have been developed to elicit the tacit knowledge of experts [136]. Their popularity appears to have continued to 2023 as Matetic et al. [105] classified all the four latest AHU knowledge discovery approaches, or knowledge-based approaches, as rule-based. For AHUs, House et al. [141] developed the AHU Performance Assessment Ruleset (APAR) that utilized the understanding of the conservation of energy and mass laws. However, these methods are limited to the experience of the domain expert. Therefore, they are not capable of detecting faults that are not flagged in historical data or those faults which are beyond the engineer's experience [132]. These rule-based algorithms are also the most popular form of FDD in the commercial setting [14], but precise configuration requirements and scope limitations are known drawbacks [8].

2.6.2 Deep Knowledge Systems

Physical models, or deep knowledge, require a greater understanding of the system for their development. The general methodology adopted is to establish a baseline of normal operation using physics-based and engineering knowledge-based models, which are then continually compared to the operating status of the AHU [19]. The differences (or residuals) between the measurements and the value of the model variables are then used to detect faults. For example, Song and Wang [142] developed a model for a cooling coil which was capable of detecting an introduced damper stuck fault. Additionally, Wang et al. [143] developed a model-based FDD strategy to diagnose abrupt faults in AHUs. Depending on the complexity of the models, these approaches may be considered as white-box or grey-box [19]. The complex models require a lot of detail to be input into simulation software. Recent work typically uses TRNSYS, SIMBAD, EnergyPlus, and OpenModelica [105]. These approaches are highly accurate but require a large number of parameters and extensive data on the physical principles governing the thermodynamic processes [105]. The less complex models are similar to shallow knowledge systems in that they have low data requirements, but they rely heavily on domain expertise and cannot account for faults that extend beyond the experience of the engineer [132]. To summarise, a deep knowledge system aims at estimating parameters and signals by applying a dynamic process.

2.6.3 Comparison

The strengths and weaknesses of knowledge-based approaches are presented in Table 2-10. Deep knowledge systems are highly accurate but their complexity is a severe shortcoming. The less detailed shallow knowledge systems do not require a complete mathematical model, therefore they are easier to integrate into complex HVAC systems and perform well in the practical setting where data is limited. However, it is also possible to effectively integrate shallow knowledge and deep knowledge systems. For example, Deshmukh et al [144] combined expert rules with first principles thermodynamic modelling to detect a stuck damper fault which led to savings of around $3400/month or 18% of the monthly cost.

Table 2-10 Summary of strengths and weaknesses of knowledge-based approaches

	Strengths	Shortcomings
Shallow Knowledge Systems	• Easy to implement and interpret • Can capture expert knowledge and guidelines • Can be effective for simple faults	• May not be able to handle complex or unusual faults • May require a large number of rules • May not be able to adapt to changing conditions
Deep Knowledge Systems	• Can handle complex and usual faults • Can adapt to changing conditions • Can provide insights into system behaviour	• Requires accurate models • May be computationally expensive • May be difficult to interpret

2.6.4 Conclusion

Shallow knowledge-based algorithms are commonly preferred to deep knowledge-based algorithms in practical applications given their lower data requirements. While they are robust to data issues, they do not offer a means of improving the dataset for data-driven approaches, in their current form. For example, the rules are generally applied at an instantaneous moment, disregarding the changing condition over time which data-driven approaches are built upon. To enhance the rule-based approach, it is necessary to expand its capabilities beyond merely detecting faults and incorporate expert knowledge for analysing degradation patterns. This offers a means of integrating engineering-based problem-solving techniques into the data-driven approach.

Another challenge experienced by rule-based systems is the diagnosis task. While House et al. [141] offer diagnostic reasoning for the occurrence of a particular triggered rule, there are many possible causes. The difficulty of distinguishing sensor faults from component faults is a known issue. However, the characteristic properties used to trigger rules may be further investigated to align the discrepancy with the fault that is known to have a similar pattern. For example, a degrading sensor may randomly switch between triggered and non-triggered states, while a degraded control valve may cause a low-level continuous fault. These are inherently different faults that would cause the rule to be triggered as they measure the same characteristic property. Therefore, the current application of expert systems through rule-based algorithms is instantaneous and high-level. This leads to uncertainty regarding the operational state of the AHU which creates an ***operational gap*** when compared to healthy AHUs that are injected with known faults in experimental or simulated studies. A means of progressing the instantaneous and high-level implementation of expert systems to a form that's more aligned to an AI/ML learning system is needed to bridge the operational gap.

The review of data-driven approaches in subsection 2.5, and knowledge-based approaches in this subsection 2.6, identified that one may address the shortcomings of the other, and vice versa. While additional effort is needed to address the data gap and operational gap, the skills gap may be addressed through the combination of both of these approaches. The integration of domain expertise with data-driven learning ability is known as a hybrid approach.

2.7 Hybrid Approaches

The combination of shallow knowledge-based or deep knowledge-based approaches with data-driven approaches may be referred to as a hybrid approach. Similarly to complex physical models, hybrid approaches may be classified as grey box approaches as they are situated between the highly interpretable white box methods and the unknown black box methods. In the combination of deep knowledge-based systems and data-driven approaches, it is common for the deep knowledge system to be used for baseline development [19]. Additionally, data-driven approaches can be used to decrease high-dimensional data to reduce the complexity of models [139]. Matetic et al. [105] provide another example whereby the supplementation of sensory data measurements could address the challenges of collecting all the relevant building information needed to build overly complex models. Including measured data is a means of increasing resilience by reducing noise, uncertainty and modelling requirements, which can lead to predictive performance that may surpass data-driven techniques [145]. The Matetic et al. [105] review found many examples of physics-based models being used to synthetically generate both healthy and faulty data to apply data-driven modelling. While the integration of knowledge (discovery) based approaches is also combined with data-driven approaches, mostly for pattern recognition and bayes-based techniques. In a generic review of bridging the gap between data-driven and model-based approaches, Tidriri et al. [146] proposed that using data-driven methods for fault detection and model-based methods for fault diagnostics has better FDD potential than using a single method. In a 2022 review of machine learning (ML) methods for building systems automated fault detection and diagnostics (AFDD), Nelson and Culp pointed out that a common conclusion in the literature is that many algorithms grow stronger when used in combination with another [8]. This suggests that hybrid approaches can leverage the benefits of both knowledge-based and data-driven methods to produce a solution that performs better than either approach would produce individually. Some examples in the literature are given in Table 2-11 with the benefits and shortcomings summarised in Table 2-12.

Table 2-11 Summary of hybrid approaches for AHU FDD in the literature (blank cells represent "not specified")

Technique	Software	Year	Ref
LSTM-SVDD	OpenModelica	2022	[127]
2-D CNN	TRNSYS	2021	[109]
MLP	TRNSYS	2021	[110]
RACNN	Not Specified	2021	[112]
Bayesian inference	Not Specified	2021	[113]
LSM	Not Specified	2020	[147]
PCA	HVACSIM+	2020	[126]
ARX, RF	IDA ICE	2020	[117]
TARM, CART	Not Specified	2020	[130]
DBN	Not Specified	2020	[116]

Table 2-12 Summary of strengths and weakness of hybrid approaches

Strengths	Shortcomings
• Can combine the strengths of both knowledge-based and data-driven approaches • Can handle a wider range of faults	• May be more complex and difficult to implement • May require significant domain knowledge

2.7.1 Application of Hybrid Approaches

As discussed in subsection 2.3.7, to implement a proactive maintenance strategy, the understanding of the domain area must be integrated with data-driven techniques. Given the ability of hybrid approaches to integrate both knowledge-based and data-driven approaches, it appears to be the most capable means of performing the analysis needed for proactive maintenance. There are many examples in the literature of one-off combinations of specific methods, of which there are many to choose from. However, as the well-known dictum in data science states, “There is no one-size-fits-all algorithm”. Therefore, for proper implementation, Tidriri et al. [146] point out the need for a generic framework.

In the Matetic et al. [105] review, the authors provide some guidelines for building an anomaly and fault detection model for HVAC systems. To aid the FDD approach selection, the authors developed the guide in Figure 2-14 to navigate the issues of lack of labelled datasets, data quantity, selecting appropriate features, the lack of research on estimating the severity of faults in component degradation, and real-world constraints regarding data collection. Given the issues discussed in subsections 2.2 - 2.7, the most likely route is shown in red on the right-hand side of Figure 2-14. Beginning at the top with obtaining the dataset, it is likely that the dataset will be insufficient due to the data gap, even if there are many records available. Given the need for a human-centric approach in the Industry 5.0 paradigm, an expert will be involved that can bring tacit knowledge along with building information to address the knowledge gap. Therefore, a knowledge-based, or knowledge-discovery approach, is most suitable. This approach may then be expanded towards the AI/ML learning systems to develop hybrid approaches if the operational gap is addressed. Other researchers appear to agree that hybrid approaches are most suitable as Matetic et al. [105] found that hybrid approaches are the most popular classification type at 47% of the reviewed literature, followed by data-driven at 36% and knowledge-based at 17% (composed of physics-based at 9% and knowledge discovery at 8%).

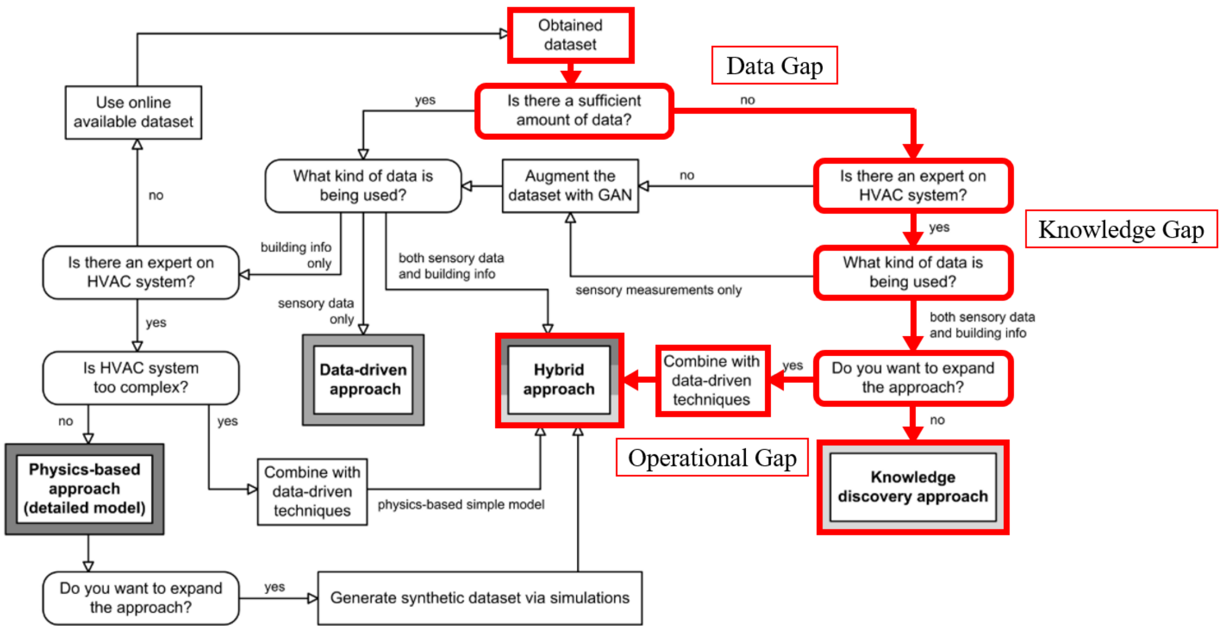


Figure 2-14Matetic et al. [97] guide to selecting an appropriate FDD approach

Once the approach is selected, the next challenge becomes selecting the appropriate modelling technique or type of data-driven approach. The guideline proposed by Matetic et al. [105] is visualised in Figure 2-15. Given the lack of data and lack of data labels in real-world industrial datasets, it is most likely that PCA techniques will not be needed and that unsupervised learning may be most suitable initially. Depending on the site-specific problems that the domain expert encounters, a clustering approach may be suited to identify abnormal data points with shared properties, or an association rule mining approach may be useful to explain relationships in the data. Over time as faults or failures begin to be recorded, supervised learning may be possible to identify repeat occurrences of past faults. Assuming the data is not heavily imbalanced, a domain expert may choose an approach that is capable of identifying multiple faults with a dataset that is likely to be of insufficient quantity as few labels are available. Therefore, a decision tree method may be most applicable. Some of the advantages and disadvantages of particular supervised and unsupervised ML used for HVAC FDD as identified by Matetic et al. [105] are summarised in Table 2-13.

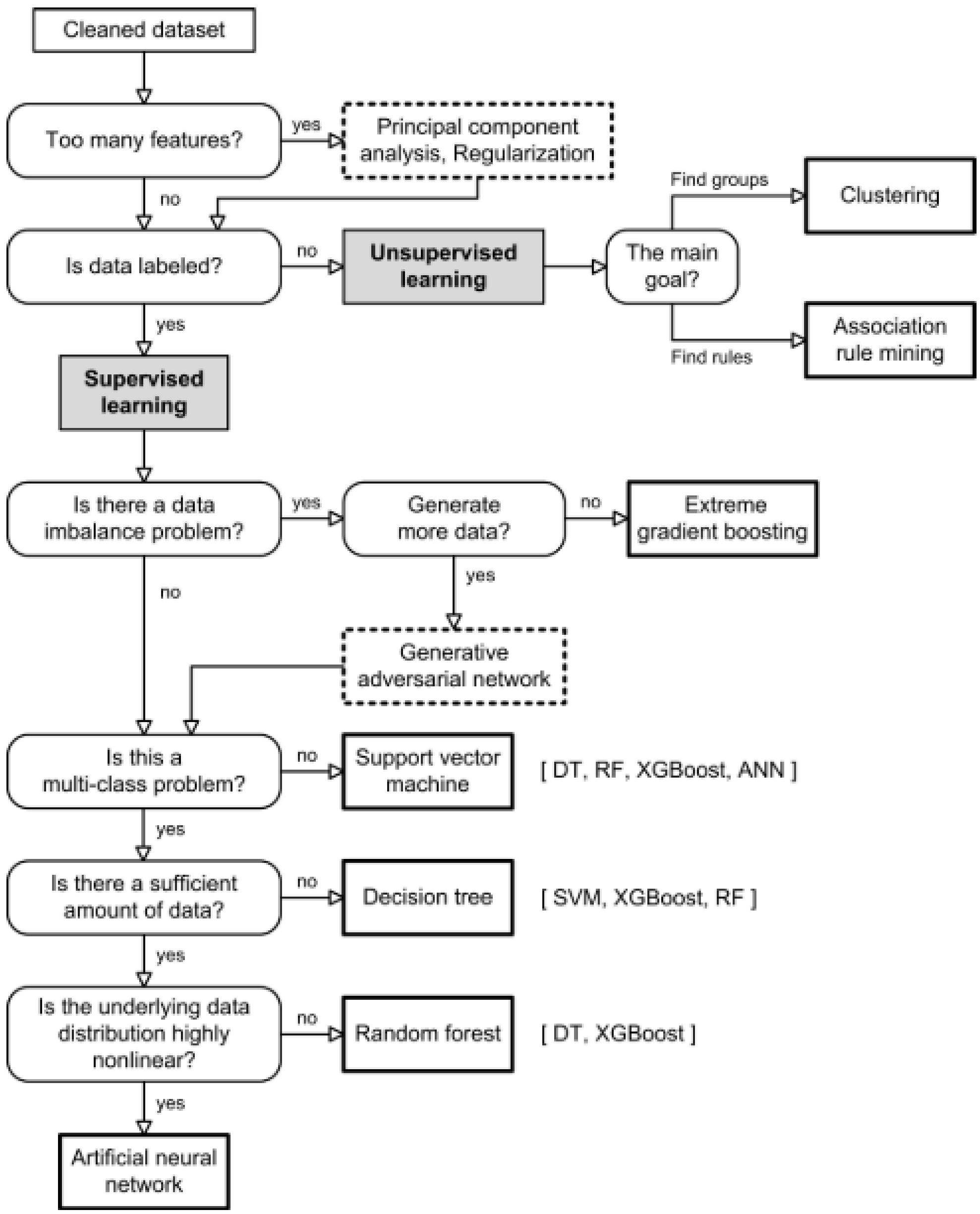


Figure 2-15 Matetic et al. [105] *guide to selecting an appropriate modelling technique*

Table 2-13Matetic et al. [105] *summary of ML models in the field of FDD for HVAC systems. Models one to 5 are supervised while the final two are unsupervised*

Model	Advantages	Disadvantages
SVM	– Ability to seamlessly expand the feature space through the use of kernels. – Modeling (non-linear) distributions with relatively high accuracy. – Overall good performance due to maximum-margin separation.	– Requires longer training times for learning from larger datasets. – Low interpretability.
ANN	– Capable of learning complex spatially- and/or temporally-dependent fault nature. – Capable of simultaneously learning both feature extraction and the objective in an end-to-end optimization process (i.e., deep learning). – Ability to diagnose multiple faults simultaneously. – Model learning can be parallelized.	– Performance highly dependent on network architecture, which is influenced by numerous hyperparameters that need to be tuned manually. – Deep architectures usually take longer to train. – Low interpretability.
DT	– Ability to model highly complex representations, through input space partitioning. – Can achieve excellent performance. – Highly interpretable.	– Model performance may not be guaranteed in complex multi-class faults. – Susceptible to overfitting.
RF	– The best trade-off between model complexity and generalization ability. – Usually is more accurate than DTs. – Robust to outliers. – Model learning can be parallelized.	– High computational complexity due to ensembling. – Low interpretability.
XGBoost	– Efficient with large datasets. – Insensitive to distribution skewness.	– Difficult to tune. – Performs poorly on sparse and unstructured data. – Low interpretability.
Clustering	– Ability to discover patterns among the data points. – Can extract useful information from intra-cluster relationships.	– More sensitive to data quality. – Similarities can be difficult to find in practice, especially when classifying multiple faults. – New faults that are not recognized as a member of created groups cannot be diagnosed properly. – Low interpretability.
ARM	– Ability to identify strong rules in the data. – Tool for finding inherent regularities in the data. – Highly interpretable. – Rule learning can be parallelized.	– Usually requires categorical variables as input. – The results often consist of a large amount of redundancy, which requires post-mining steps. – The complex connections between multiple features can be challenging to interpret in comparison to supervised methods.

2.7.2 Conclusion

While these guides are very useful, there is still a lack of a structured means of aiding the transition to proactive maintenance given the issues that are caused by the knowledge gap, data gap and operational gap. In the literature, a myriad of data-driven approaches have been developed and there is an increasing trend [18]. However, most of the commercial tools (12 out of 14) assessed by Granderson et al. [14] use rule-based algorithms. The disconnect between academia and commercial tool providers suggests that digitalisation-based techniques are not yet a viable option for scaled deployment. Hence there is a lack of real-world case studies of deployed solutions in the industrial setting in the literature for other researchers to follow. This is a ***practice-theory gap***.

2.8 Main Barriers to the Application of Proactive Maintenance on AHUs in the Industrial Setting

The following subsection, 2.8, synthesises the reviewed literature to uncover the main barriers to proactive maintenance of AHUs in the industrial setting. As discussed in subsection 2.1.1, the Industry 5.0 objectives guide European research activity to be human-centric, resilient and sustainable. The four research gaps identified in this chapter (knowledge, data, operational and practice-theory) create the chasm between the state-of-the-art and the Industry 5.0 paradigm. Therefore, in Table 2-14, each of the four identified research gaps is aligned to the relevant Industry 5.0 objective they inhibit. Furthermore, each research gap is accompanied by a research barrier. To define the research barriers a delineation used in a 2022 review of computing-based automated FDD of HVAC systems [19] is used as inspiration. In this review, Chen et al. [19] summarised the state of the art of FDD development from both a data perspective and an approach perspective. The issues with real-world data are a clear barrier that causes the data gap. But the other gaps are not being addressed because the approach used is either not adequate or there is no approach readily available. This is explained further in the following subsections where the barrier to each research gap in Table 2-14 is discussed.

Additionally, potential solutions to overcome the identified barriers are discussed, taking into consideration the data analysis levels outlined in Table 2-4. The lack of domain guided-analysis is a barrier causing the knowledge gap which suggests a solution is required at the conceptual level to integrate and guide the problem-solving abilities of engineering and data science. As this type of solution will require several steps and process models, an overarching term is needed to describe the overall process. As discussed by Brem [148], and presented in the GreenSAM project [149], a framework is an appropriate term for this type of solution. The framework may contain multiple methodologies and acts as an umbrella term used to describe the general approach taken to achieve an aim [149]. Therefore a framework is needed to provide structure for performing goal-oriented analysis and exploration-based analysis. Secondly, the data issues causing the data gap must be identified and addressed. A dedicated phase in the framework is needed for the identification of a variety of data issues that may be developed into a specific method for detecting data issues in AHUs. Thirdly, the uncertainty regarding the operation of AHUs in the industrial setting requires an assessment using a knowledge-based approach to create a readiness indicator for data-driven analysis. A dedicated phase in the framework is needed for this task. Lastly, the reproducibility and practicality of the analysis may be tested through the deployment of the framework at the algorithm and tool level. The gap between theory and practice may be bridged by documenting the process of translating the conceptual framework to a specific set of instructions for the development of a software program.

Table 2-14 Research objectives, gaps, barriers, questions and potential areas of novelty

Industry 5.0 Objective	**Gap**	**Barrier**	**Simplified Research Question**	**Novelty Potential**
To be human-centric	Knowledge Gap	Lack of domain-guided analysis	How can engineering and data science problem-solving techniques be integrated?	Development of a framework to aid the application of a hybrid approach
To be resilient	Data Gap	Data Issues	How can an industrial dataset be curated and assessed?	Implement the framework on real industrial AHU data to develop a dataset ready for analysis
To be sustainable	Operational Gap	Lack of readiness indicator	How could the faults limiting the application of data-driven techniques be quantitatively monitored?	Implement the framework on the curated dataset to identify sensor, component and control faults
All three	Practice-theory Gap	Lack of Reproducibility & Practicality	What are the practical considerations to be addressed in tool deployment?	Identification of practical considerations during the deployment of a tool for proactive maintenance

2.8.1 Lack of Domain-Guided Analysis

The first Industry 5.0 objective is to be human-centric, which is to involve people. This means that the knowledge obtained by an engineer from experience, and the knowledge obtained by a data scientist from the data, must be integrated effectively. For the initial implementation of data-driven analysis, domain knowledge will be required to contextualise the dataset, especially considering the lack of labelled data. For this reason, knowledge discovery will need to occur to understand the dataset and potential fault patterns. Unsupervised methods are capable of discovering this knowledge but most of the knowledge discovered is useless [17]. For example, an unsupervised algorithm may detect a large number of clusters in a dataset due to the differing modes of operation of an AHU. While this may lead to some interesting comparisons between heating and cooling modes, the insight obtained would likely not lead to any energy-saving actions as the AHU may be operating as designed and not in need of any maintenance. Therefore, from an approach perspective, a domain knowledge-guided data mining approach is key to maximising the effectiveness of the exploratory task. How the integration of these two distinct disciplines is achieved needs further investigation.

In the Industrial AI approach outlined in subsection 2.4, the majority of manufacturing problems can be classified as either visible problems or invisible problems [150], as in Figure 2-16. Invisible problems encompass “machine degradation, lack of lubrication, loss of accuracy, wear and tear of parts, and resource waste” [150], while visible problems are often the result of invisible factors such as “component failure, equipment downtime, machine break-downs, decrease in product quality” [150]. These problems need to be both solved and avoided, requiring different approaches. The visible problems are familiar to domain experts and goal-oriented CRISP-DM solutions are suitable, especially in the Visible and Avoid quadrant [150]. The invisible problem quadrants will require an exploratory approach as the problem is less defined. Based on these differing trajectories, the DST map was proposed by Martínez-Plumed et al. [87]. The outer circle, in Figure 2-11, contains the exploratory activities and the inner circle contains CRISP-DM or goal-oriented activities. For industry to transition from reactive to proactive approaches, goal-oriented CRISP-DM visible problem-solving and avoidance will need to evolve to data-oriented exploratory invisible problem-solving and avoidance. That is to say, the visible problems must be adequately managed before the value of data-driven analysis may be realized through invisible problem-solving. While originating from the two distinct disciplines of engineering and computer science, an integrated approach would fill a knowledge gap, enabling a seamless transition from reactive maintenance to proactive maintenance practices in the era of digitalisation. Many integration-type adaptation studies have shown that CRISP-DM may be successfully combined with domain methodologies [94]; however, evolving from visible problem-solving to invisible problem-solving requires further study.

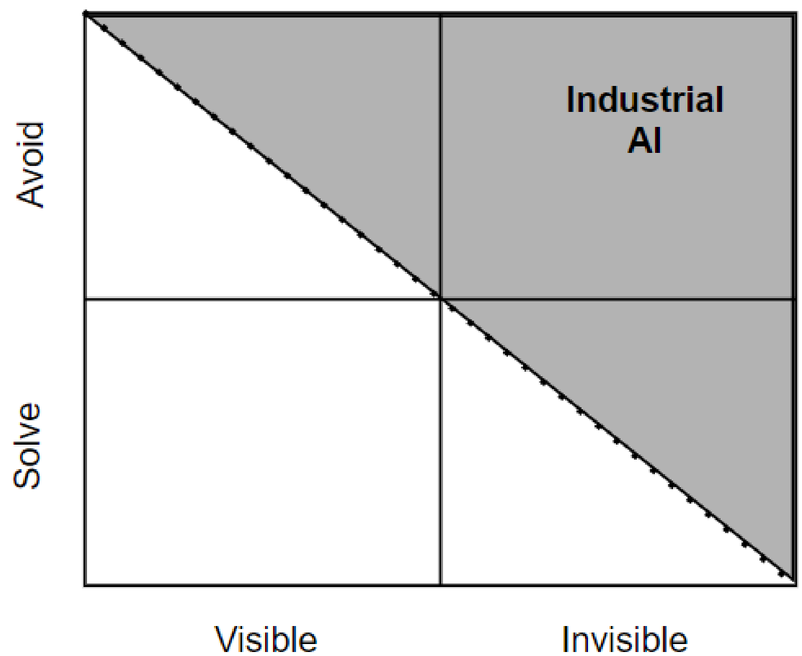


Figure 2-16 The four quadrants of industrial AI opportunity space based on the study of Lee et al. [150].

2.8.2 Data Issues

To be resilient, the solution must be applicable given the variety of real-world data issues. While real-world buildings may have collected a large quantity of operational data, these datasets are generally unlabelled and not ready for advanced analytics such as prediction tasks [128]. Furthermore, industrial facilities may not have large amounts of data as 62% of facility managers say they are deficient in terms of data collection and analysis for their

HVAC systems [58]. Statistical techniques such as the adjacent information recovery (AIR) filter [119] and generative adversarial networks (GANs) [118] have been used to address incomplete datasets in AHU FDD but the severity of the data issues in industrial facilities needs further research. Therefore, a means of systematically curating a dataset ready for both assessment and analysis is needed to address the data issues barrier.

While the visible and invisible problem-solving concept provides direction to fill the knowledge gap, further direction is needed to fill the data gap. In general, the current state of the data in the industrial setting appears to be insufficient for proactively maintaining AHUs, based on the survey of hi-tech facility managers [58]. Therefore, there is a need to identify issues and get them fixed. While verifying data quality is a task within the data understanding phase of CRISP-DM [88], the severity of the data issues regarding AHUs in the industrial setting requires greater significance to be placed on the activity of identifying issues. To summarize the data issues identified in the literature, Table 2-15 categorizes the main data-related challenges outlined by systematic reviews of data-driven applications in both generic industrial and HVAC-specific studies. The table is by no means exhaustive but does indicate that many challenges are pervasive in both fields.

Table 2-15 Classification and comparison of the issues identified in systematic reviews in the area of advanced data analysis in the industrial domain. Note: 'X' indicates the challenge is mentioned.

Classification	Challenge	General Industrial				HVAC-Specific	
		Dogan and Birant [22]	**Bertolini et al. [23]**	**Wuest et al. [24]**	**Dalzochio et al. [25]**	**Mirnaghi and Haghighat [26]**	**Zhao et al. [18]**
Broken	Data Availability	X	X	X	X	X	
Bad-Quality	Noisy Data	X				X	
Background	Algorithm Selection	X	X	X	X		
	Evaluation of Results	X	X	X	X		
	Distinguishing between sensor and component faults						X

In Table 2-15, the data issues are classified using the categories proposed in the Industrial AI paradigm; broken, bad-quality and background, which are collectively known as the "3B's" [80]. Firstly, data may be considered broken if it does not fulfil the requirements of the analysis. Professor Lee [80] outlines that "data can't just be numerous—it must be comprehensive". As in Table 2-15, data availability is a common issue in the literature, whereby the data is described as missing, incomplete or insufficient. Data may be unavailable for reasons such as a lack of sensors, faults in communication protocols or damaged sensor networks. In the case where the necessary information is available, it must accurately represent the physical phenomena it measures. Bad-quality data does not reach the standard required to apply data-driven techniques and is often considered noisy. Noisy data is composed of issues such as incorrect data, improper data, duplicated data and inconsistent data, which is often caused by sensor faults. The different fault types are visualized in Figure 2-17 and are commonly classified as drift fault (a), bias fault (b), precision degradation fault (c), spike fault (d) and stuck fault (e) [151]. The presence of a sensor fault may lead a data-driven algorithm to incorrectly classify the anomaly as a component fault, which would lead to a lack of trust in these methods by industrial practitioners. Distinguishing between sensor faults and component faults is a key challenge in HVAC applications [18], which requires an understanding of the operation of the equipment to overcome. This is a key component of the background issue that relates to the need for understanding the context of the data. In practical applications, the data is often inadequately labelled, which means the interpretation of domain experts is needed to clarify the meaning of each measurement [152]. Uncertainty about the meaning behind the data leads to the prevalent challenges of selecting an appropriate algorithm and evaluating its results. If the meaning behind the data is not well understood, the value of data-driven analysis diminishes as the results cannot be adequately interpreted.

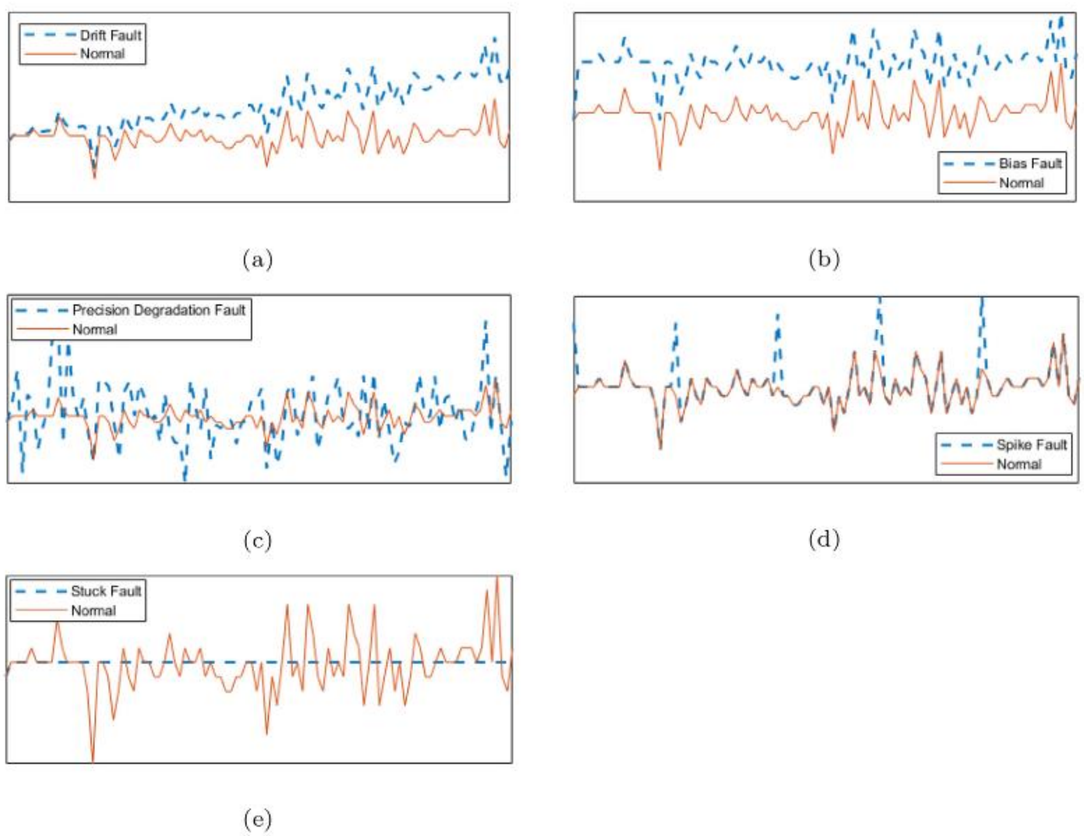


Figure 2-17 Illustrations of (a) drift, (b) bias, (c) precision degradation, (d) spike, and (e) stuck faults. Reprinted from Jan et al. [28], with permission from Elsevier, 2020).

The “3B” data issue taxonomy provides a classification for the plethora of data-related challenges in the industrial setting that extends beyond the challenges outlined in Table 2-15. The nature of the data issues is interpreted as shown in Figure 2-18, where data challenges can be classified into three regions in the Venn diagram. For example, the challenge of distinguishing sensor faults from component faults is a key challenge in HVAC applications [18], which may be caused by faulty sensors (Bad-quality) or the physical failure of mechanical components (Background). Therefore, rather than prescriptively resolving a myriad of specific data challenges, a means of managing the challenges that lie within the Venn diagram is a potentially more sustainable solution. Therefore, the “3B’s” classification provides an appropriate concept to deal with the inadequacy of real-world industrial data as it aligns with developing resilient solutions that are necessary for a successful digital transition.

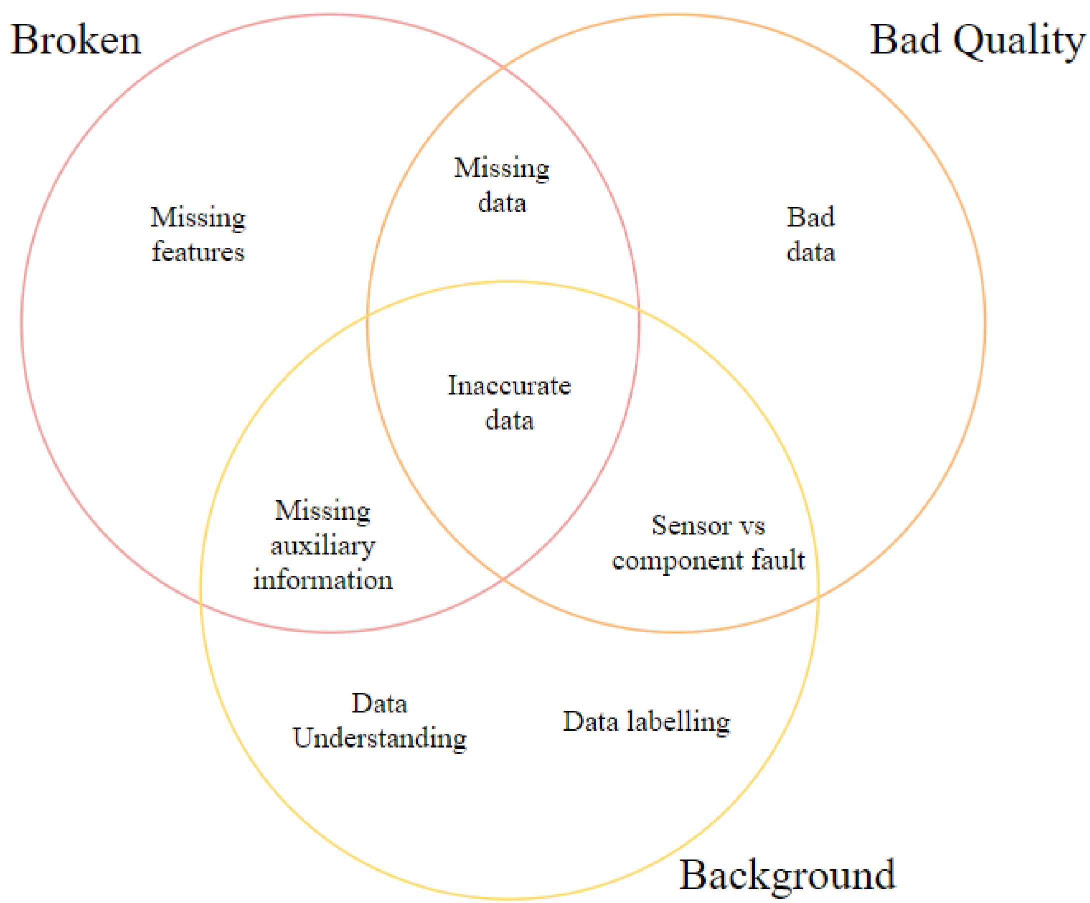


Figure 2-18 Data Issue Map

2.8.3 Lack of Readiness Indicator

In real building applications, the labelled data needed for supervised methods are very hard to obtain [153] and the high-quality data needed for both supervised and unsupervised techniques are not readily available [58]. While the data issues are a significant barrier, there is also the problem of determining when a particular fault has taken place as discussed in subsection 2.6.4. Data-driven approaches typically involve three steps [154]: establish a baseline for fault-free status, compare incoming data to the baseline, and flag a fault based on the difference. This creates uncertainty regarding the point at which data-driven approaches become applicable as it is difficult to determine what a "fault-free" AHU looks like. An indicator or similar is required to determine when the data-driven analysis is feasible, but there is a lack of an approach to follow for this purpose. This indicator is likely to depend on the application area and will therefore require domain expertise to determine appropriate thresholds or conditions to be achieved. For example, the majority of FDD methods require steady-state conditions for their application as it is difficult to model and predict transient conditions [155]. Therefore, a means of monitoring issues such as transient states is required. This understanding of the operation, or the operational gap between real-world and theoretical performance, would improve the usefulness of the analysis for increasing the sustainability of the unit.

In subsection 2.6.4, the difference in operation between real-world AHUs in the industrial setting and simulated or experimental AHUs in the literature are discussed and the extension of rule-based approaches was proposed as a potential avenue for further research. A phase dedicated to the extension or modification of existing expert systems for the task of monitoring the degradation pattern would provide a means of measuring the readiness for data-driven approaches. A method that assesses the visible faults, those understood in the engineering domain with a measurable fault characteristic, would provide a quantitative indicator for the readiness for invisible problem-solving. Without this indicator, visible faults may be active in the system which would compromise the ability of the data-driven approach to detect the invisible faults.

2.8.4 Lack of Reproducibility & Practicality

While there is an immense number of different algorithms applied to building systems FDD [83], the majority of publications focus on increasing the fault detection accuracy of their chosen method on simulated or experimental datasets. But this limits the applicability of FDD methods in the real-world as stated by Melgaard et al. [83], "repetitive use of the same datasets, combined with a focus on theoretical research, can impose a challenge in the actual implementation of FDD". While guides exist to aid method and approach selection [105], there are many more steps required to transition from experienced-based maintenance decision-making to data-driven maintenance decision-making. Therefore, solely focusing on FDD is not the correct approach. A pilot analysis or case study is required to tackle practical issues and aid the reproducibility of the analysis [83]. The lack of a means of systematically developing reproducible and practical analysis is a barrier that must be overcome.

In the industrial sector, decision support tools often leverage the concept of a digital twin. The literature on digital twins for maintenance was reviewed in 2020 by Errandonea et al. [156]. In the review, distinctions are made based on integration modes between the physical asset and the digital component as identified by Kritzinger et al. [157]. The different types are digital model, digital shadow and digital twin as shown in Figure 2-19. At the first level, data is manually fed between the asset (physical object) and the digital object. This may be categorised as offline analysis that requires manual effort to obtain the data and relay the results to action maintenance activity. At the second level, data is automatically received from the physical object and analysed by the digital object. This may be considered as online analysis as the results are automatically presented for interpretation by an engineer to then schedule maintenance activity. The final level involves automatically feeding the results from the digital object back to the physical object. Given the need for human-centric solutions and the scope of this research being limited to observational-based analysis, a digital twin is not suitable. Therefore, the development of a digital shadow-type tool is the appropriate means of deploying the analysis.

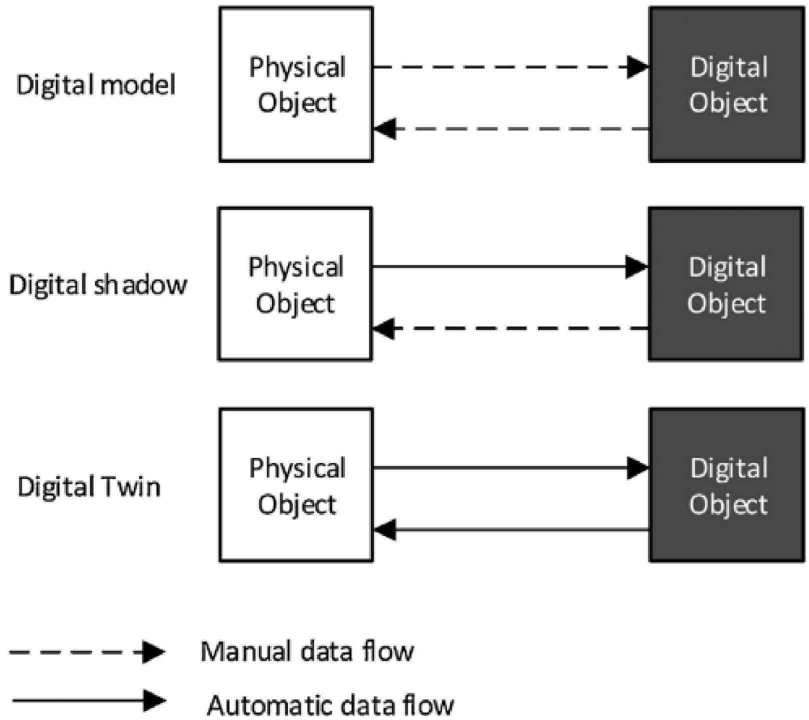


Figure 2-19 Integration modes in the Digital Twin concept

2.9 Conclusion & Next Steps

This chapter has reviewed the state of the art in the area of proactive maintenance for AHUs in the industrial setting. Four key research gaps have been identified;

Firstly, the knowledge gap is the difference between the reactive problem-solving ability of those in the engineering domain and the proactive problem-solving ability of those in the data science domain. The lack of a domain-guided data mining approach is a barrier that must be overcome to address this gap. The research question this raises is *"How can engineering and data science problem-solving techniques be integrated?"* as in Table 2-14.

Secondly, the data gap is the difference between the high-quality datasets obtained from simulated and experimental studies in the literature and the poor-quality datasets obtained from the industrial setting. The issues with real-world datasets cause this gap and there is a lack of publicly available industrial datasets to encourage research on this key problem. This raises the question *"How can an industrial dataset be curated and assessed?"*

Thirdly, the operational gap is the difference between the known operating state of AHUs in the simulated and experimental studies in the literature, and the unknown operating state of AHUs in the industrial setting. The lack of a readiness indicator to determine when data-driven approaches are viable is a barrier. This leads to the question of *"How could the faults limiting the application of data-driven techniques be quantitatively assessed?"*

Lastly, the practice-theory gap is the difference between the predominantly data-driven approaches in the literature and the predominantly rule-based approaches in the commercial and industrial settings. The lack of reproducible and practical approaches in the literature is a barrier causing this gap. This poses the question *"What are the practical considerations to be addressed in tool deployment?"*

While these are four distinct gaps, they are not mutually exclusive and do contribute to one another's occurrence. For example, in a 2022 article titled *"Does a knowledge gap contribute to the performance gap? Interviews with building operators to identify how data-driven insights are interpreted"*, Markus et al. [158] exposed potential barriers to widespread adoption of data-driven approaches in industry. One such barrier is the unfamiliarity building operators and facility managers may have with data-driven methodologies and tools. The authors note the inherently risk-averse nature of building operators which leads to hesitancy regarding the use of data-driven approaches. Therefore, the knowledge gap appears to be the main threat to the adoption of data-driven approaches. The development of a research methodology to address this gap, and the other gaps, is discussed in Chapter 4.

Table 2-16 Literature Review Overview

Problem	Gap	Barrier	Objective	Potential Novelty	Output	Chapter
Reproducible & usable data-driven analysis for the proactive maintenance of AHUs in the industrial setting	Knowledge	Lack of domain-guided analysis	Integration-type adaptation of CRISP-DM	Visible and invisible problem-solving for AHU faults	Journal Article 1	4
	Data	Data issues	Dataset curation	Industrial AHU dataset	Journal Article 2	5
			Data assessment method	Domain knowledge integrated data assessment	Journal Article 3	5
	Operational	Lack of a readiness indicator	Expert system modification (ie APAR)	Novel application of existing methods	Journal Article 3	5
	Practice-Theory	Reproducibility & Practicality Issues	Deployment of a digital shadow	Identification of novel practical considerations	Journal Article 4	6

3 CHAPTER 3 – METHODOLOGY

3.1 Introduction

The literature review in Chapter 2 led to the discovery that further guidance is needed to implement proactive maintenance on AHUs in the industrial setting. To address the research questions and overcome the gaps and barriers in Table 2-14, it is necessary to develop research that guides the development of proactive maintenance, from conceptualisation to implementation. For example, the knowledge gap is a theoretical problem that requires a structured approach to integrate distinct philosophies, concepts and methodologies. The data and operational gaps also require a structured approach but both of these gaps are more defined than the knowledge gap so a more specific approach may be taken to close these gaps. Lastly, the practice-theory gap requires actual application in the field to bridge the gap between academia and the industrial setting. This gap should therefore be addressed by software that delivers a specific task, such as the development of a tool.

Hence, the research methodology should commence by constructing a framework that tackles the existing knowledge gap. This framework should integrate process models to address the data and operational gaps at the method and algorithm levels. Subsequently, the research can progress towards addressing the practice-theory gap at the tool level. The effectiveness of this research will be measured by its ability to provide a solution that facilitates proactive maintenance decisions and reduces energy usage in a real-world case study site.

3.2 Proposed Research Methodology

While research often prioritises data modelling, it’s the remaining stages of data analysis, such as data understanding, preparation and exploratory analysis, that need further attention [159]. Similar to the data modelling phase, there are numerous approaches or methods available to carry out each task. Given the need for further guidance in this research area, a research methodology is proposed in Figure 3-1 to aid the development of practical and reproducible analysis.

In Step 1 of Figure 3-1, a framework is developed to provide the general concept and structure to achieve proactive maintenance tasks. The framework is comprised of CRISP-DM and DST-map process models that have a series of steps or stages to achieve goal-based and exploratory-based analysis, respectively. In Step 2, the methods or specific techniques that may be used to address the data and operational gaps are developed. For example, a dataset may be systematically curated and then assessed. The specific algorithm used to carry out these tasks is then implemented at the algorithm level. Finally in Step 3, a tool may be deployed using a software program that delivers the data analysis to the user.

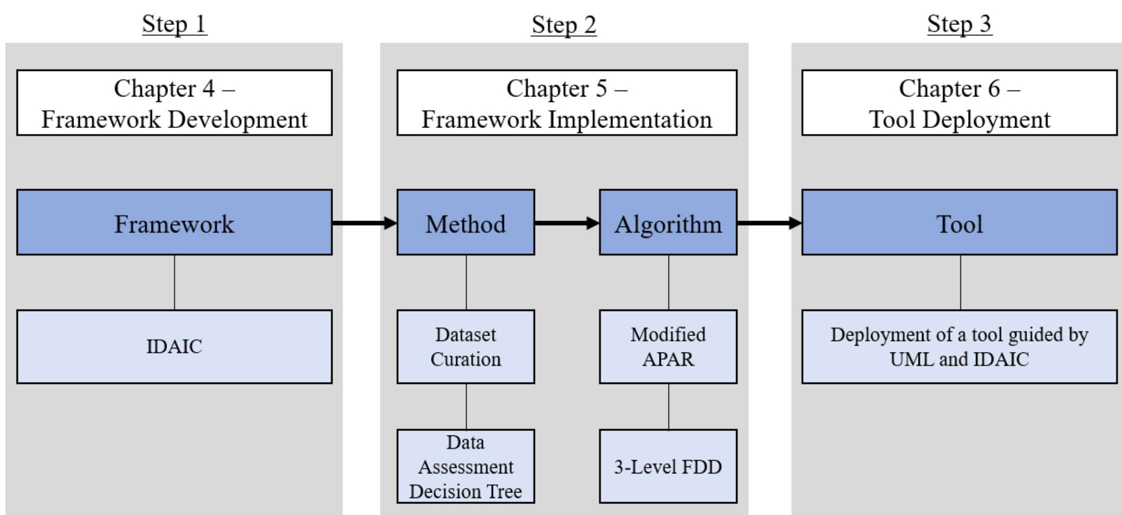


Figure 3-1 Research methodology and corresponding thesis chapter

The review of the literature in Chapter 2 uncovered the need for a novel approach to the proactive maintenance of AHUs in the industrial setting. In this chapter, a three-step research methodology is proposed to step through the full process from the development of an overarching framework to the deployment of a practical tool. Therefore, the upcoming chapters are aligned with this research methodology and structured as follows:

- Chapter 4 discusses the development of a framework that aids the transition to the proactive maintenance of AHUs in the industrial setting by integrating methodologies and process models such as CRISP-DM and DST-Map.

- Chapter 5 outlines the implementation of this framework in an industrial case study site which develops methods and algorithms to aid the identification of data issues and energy-consuming faults.
- Chapter 6 discusses the deployment of the framework as a tool for continuous improvement and proactive maintenance of the AHU in the case study site.

4 CHAPTER 4 – FRAMEWORK DEVELOPMENT

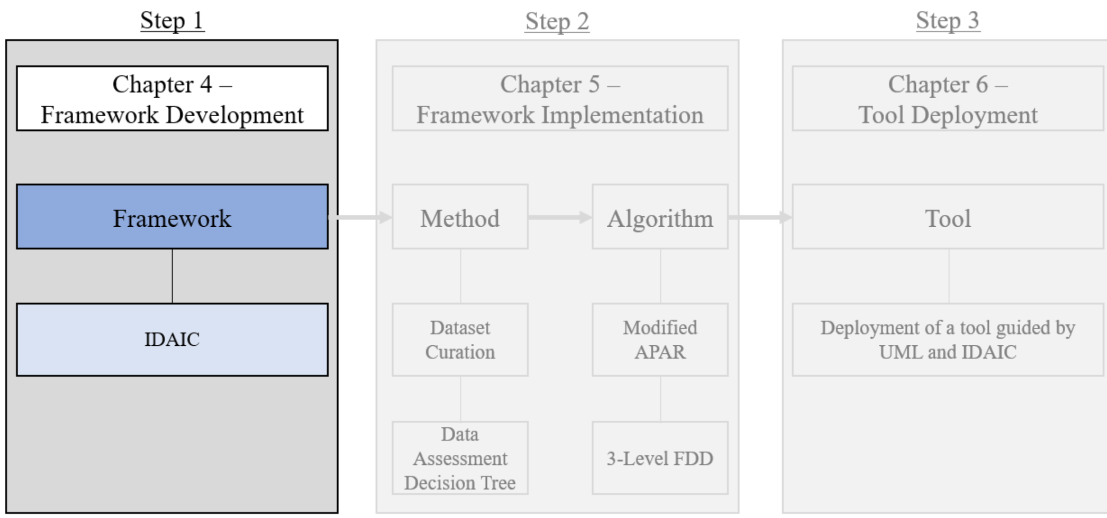


Figure 4-1 Step 1 of the research methodology

4.1 Introduction

The research objective of this chapter is to address the knowledge gap by developing a framework that enables the transition to the proactive maintenance of AHUs in the industrial setting. This would address the first research question, "*What distinguishes engineering and data science in terms of problem-solving abilities, and how can the strengths of both disciplines be integrated to develop a proactive maintenance solution for AHUs in the industrial setting?*". In Chapter 2 the differences between engineering-based and data science-based problem-solving were discussed. This led to the concept of using knowledge-based approaches for visible problem-solving and data-driven approaches for invisible problem-solving. As there is a lack of domain-guided analysis in the literature for this purpose, the development of a framework that integrates these two approaches to aid the transition to proactive maintenance is outlined in the following subsections.

4.1.1 Background

Visible problems in AHUs are well understood by domain experts as they solve and subsequently avoid these problems daily. These types of problems typically lead to a loss of function, such as failing to achieve the supply air temperature set point. In this scenario, the thermal discomfort experienced by the occupants would lead to a complaint, causing maintenance activity to take place. However, there are problems that occur which are not as easily detected. As AHUs are self-correcting in nature, they can compensate for faults so that the overall function may still be achieved. One example would be increasing the heating supplied to rectify unwanted cooling from a leaking cooling coil valve. These invisible problems are a significant contributor to wasteful energy use in industrial facilities as unnoticed faults may account for up to 20% of the energy consumed by these systems [9]. Additionally, it is estimated that 25-50% of energy wasted from faults in UK buildings HVAC systems could be reduced to 15% if faults were detected and identified early [160]. Therefore, the knowledge discovery ability of data-driven techniques is needed to identify these invisible or unnoticed faults, so that maintenance activity may be scheduled to remediate them. However, given the complex nature of manufacturing [152], data analysts do not generally understand the operation of AHUs which are often used to maintain pressure profiles or relative humidity levels as well as provide thermal comfort and adequate ventilation needs. Meanwhile, domain experts do not generally fully understand the techniques capable of uncovering invisible faults. This is referred to as the "knowledge gap".

It would appear that this is a difficult gap to bridge as 40% of hi-tech facility managers admit to remaining in reactive mode for their HVAC maintenance strategy in a 2022 survey [58]. The fundamental differences between reactive and proactive maintenance strategies were discussed in subsection 2.2, with the ability to learn from data being a key differentiator. However, if the data is deficient, which 62% of facility managers in the same 2022 survey say it is [58], then a purely data-driven learning system is not applicable. Therefore, expert knowledge and data-driven learning must be integrated. This also satisfies the need for human-centric approaches which will be more prevalent as Industry 5.0 becomes ubiquitous.

As current reactive strategies are not progressing the field towards proactive approaches given their low data requirements, a solution that increases the emphasis on obtaining high-quality data while also facilitating reactive strategies is needed. The review of the data analysis landscape in Chapter 2 identified CRISP-DM as the de facto standard for data mining, capable of being successfully adapted to integrate domain methodologies. In the following subsection 4.2, the shortcomings of CRISP-DM in relation to conducting analysis for proactively maintaining AHUs are discussed. The framework development process which addresses these CRISP-DM shortcomings is then outlined in subsection 4.3. The novel framework developed and proposed to address the knowledge gap is then detailed in subsection 4.4 with the novel contributions highlighted in subsection 4.5.

4.2 Shortcomings of CRISP-DM for proactive maintenance

Invisible problems often lead to visible problems [150]. Therefore, invisible problem-solving enables proactive decision-making to take place. By their nature, invisible problems are not easily identifiable and advanced techniques, beyond human comprehension, are required to reveal them [80]. While CRISP-DM is not a process model for proactive maintenance, it is the de facto standard for data mining and an appropriate foundation to build a framework with invisible problem-solving ability. Furthermore, this framework must integrate domain expertise to solve visible problems, integrate AHU commissioning to physically carry out maintenance activity and integrate an appropriate means of addressing data issues. However, CRISP-DM is a blueprint for data mining and was not conceived to carry out these functions. Given the literature review in Chapter 2 elicited the existence of these gaps, it would appear that CRISP-DM does not provide sufficient guidance or direction on the necessary tasks to close these gaps. Therefore the shortcomings of CRISP-DM in relation to developing a proactive maintenance solution for AHUs in the industrial setting are; 1) the lack of domain knowledge integration, 2) the lack of integration with existing strategies, 3) insufficient data quality monitoring, and 4) the lack of an exploratory analysis guide.

4.2.1 Lack of domain knowledge integration

While the business understanding phase of CRISP-DM aims to guide a data analyst through the process of understanding the business or domain, the complexity of the industrial setting makes this task difficult, hence the occurrence of the knowledge gap. It is pivotal that the problem understanding is elicited early in the analysis to avoid expending “a great deal of effort producing the right answers to the wrong questions”

[161]. There is an opportunity to expand this phase by focusing specifically on the necessary elements for developing a proactive maintenance solution, rather than adhering to the general-purpose constraints of CRISP-DM. In practical terms, the data analyst is unlikely to comprehend sometimes complex and site-specific AHU operations. This would lead to a time-consuming process in the business understanding phase which puts the onus on the data analyst to quickly get to the crux of the situation and then flesh out the details [161]. Therefore, integrating domain expertise would make this process more efficient. The responsibility should therefore shift from a data analyst developing a model to a data analyst and domain expert developing a proactive maintenance solution. Therefore, further phases are needed that allow the domain expertise to be integrated into the analysis.

4.2.2 Lack of integration with existing strategies

Similarly, CRISP-DM is a data mining process model that's designed to extract insight from data. In the development of a proactive maintenance solution, this is a key component to detect the early signs of degradation and the time a fault occurs. However, it's the physical act of maintenance that is critical to realise the potential savings identified from the data analysis. As in the literature review in Chapter 2, the proactive maintenance approach could include the detection of a deviation from normal operation (CBM), prediction of an imminent failure (PdM) or recommendations to increase the operating life of the AHU (PrM). These strategies are part of the Industry 5.0 vision that improves upon reactive approaches such as fail and fix (CM) and time-based (PM) maintenance. Therefore, on this transition path, CRISP-DM must be extended to have the capability to integrate with each of these existing processes. A key means of effectively achieving this integration would be to include the act of commissioning as it has multiple forms to deal with each of the strategies such as retro-commissioning and ongoing or monitoring-based commissioning.

4.2.3 Insufficient data quality monitoring

Further supplementation of the CRISP-DM process model could be made regarding the assessment of the data. The pervasiveness of the data issues in the industrial setting has been well documented in Chapter 2 and a means of detecting and rectifying these issues is needed. Therefore, the assessment of the data should be a focal point of the analysis. However, "Verify Data Quality" is only a task within the Data Understanding phase of CRISP-DM [85]. To analyse AHU data from the industrial setting, the importance placed on data assessment in the CRISP-DM process model is insufficient and deserving of a standalone phase. Further guidance is required to detect the "3B's" of industrial data (broken, bad-quality and background issues), with guidance also provided in terms of rectifying these issues.

4.2.4 Lack of exploratory-based analysis

As discussed by Martinez-Plumed et al. [87], CRISP-DM is over 20 years old and the field of data science has seen a shift from goal-oriented data mining to exploratory-based data science in that time. Hence, while CRISP-DM encompasses the essential phases for analysing known faults or addressing visible problem-solving, it lacks the necessary stages to conduct exploratory analysis of unnoticed, invisible faults. These undetected faults are responsible for up to 20% of the energy wastage in AHUs [9]. A dedicated analysis track must therefore be combined to enable the invisible problem-solving activity that is required by proactive maintenance approaches. The DST-map [87] was developed for this purpose in the generic case which may now be used as a foundation for the novel purpose of invisible problem-solving in AHUs.

4.3 Framework Development Process

Given the need to integrate concepts and methodologies from distinct fields, an underlying conceptual structure is needed. Therefore, a framework is the most appropriate type of solution. A process to develop a framework that maintains the necessary phases of CRISP-DM while also making the necessary adaptations to address the shortcomings in subsection 4.2 is needed. Therefore, the process followed by Corrales et al. [162] and Almutiry et al. [163] is adopted as frameworks, which prioritise data quality as a key component, were successfully developed by them using this approach. The tasks as outlined by this method to develop the framework are gathering, filtering and mapping, and clustering as shown in Figure 4-2. In the gathering phase, the necessary activities to be included in the framework are collected. In the following phase, filtering and mapping, the activities gathered are grouped where necessary to distil these activities into phases. In the last phase, the phases are further clustered to achieve the high-level goals of the framework.

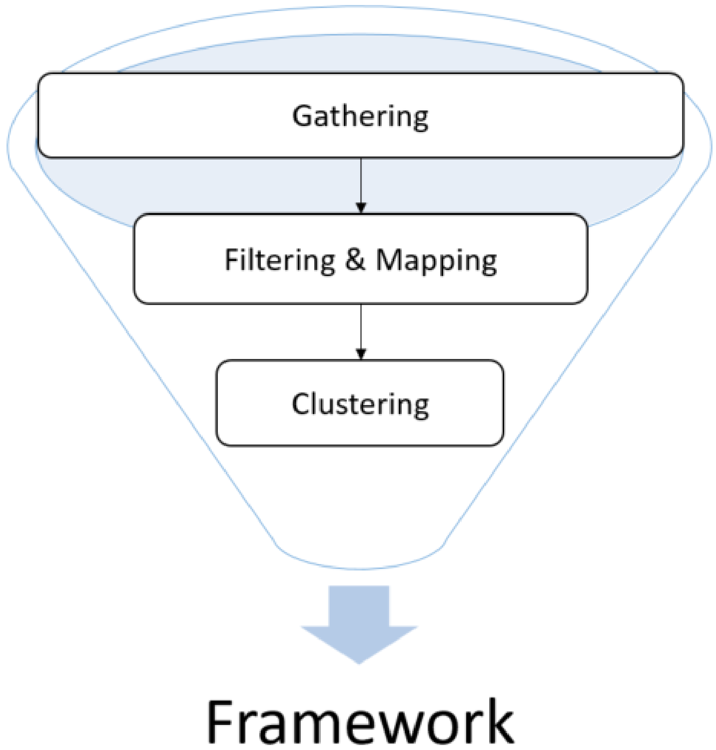


Figure 4-2 Framework development process followed by Corrales et al. [162] *and Almutiry et al.* [163]

4.3.1 Phase 1 - Gathering

CRISP-DM has four main shortcomings regarding its use to proactively maintain AHUs as outlined in subsection 4.2. Therefore, key activities, named A-I, are proposed to address these shortcomings as outlined in Table 4-1. For example, to address the lack of domain knowledge integration, an activity is proposed to elicit complex, site-specific information using domain expertise (A). Along with the exploration of the specific problems that need to be solved on the industrial site (D), these activities share the responsibility of the analysis development between data analysts and domain experts. This helps to address the knowledge gap. Meanwhile, the principles and phases of CRISP-DM are also necessary components of data analysis. Therefore, the CRISP-DM activities must be gathered in this section also and are numbered 1-7 in Figure 4-3. The activities of each CRISP-DM phase are defined based on the comparative study of KDD, SEMMA and CRISP-DM conducted by Azevedo and Santos [96]. Therefore the data understanding phase is split in two to describe the activity of acquiring the relevant data (2) and the activity of searching for unanticipated trends and anomalies (3). Each of the necessary CRISP-DM activities (1-7) and the activities to address the shortcomings (A-I) are gathered together as in Figure 4-3.

Table 4-1 Proposed activities to address the CRISP-DM shortcomings

CRISP-DM Shortcoming	Activities to address the identified shortcomings	Identifier
Lack of domain knowledge integration	Elicit complex, site-specific information using domain expertise	A
	Assess the quality and hence the value of the data using domain expertise	B
	Assess the operation of the AHU using domain expertise	C
	Explore problem areas using domain expertise	D
Lack of integration with existing strategies	Integrate the physical act of carrying out maintenance	E
Insufficient data quality monitoring	Assess the completeness of the data (i.e. detect broken data)	F
	Assess the accuracy of the data (i.e. detect bad-quality data)	G
	Assess the consistency of the data in relation to expectations (i.e. detect background issues)	H
Lack of exploratory-based analysis	Integrate exploratory data analysis activities	I

4.3.2 Phase 2 - Filtering and Mapping

In phase 2, filtering and mapping, similar activities are grouped. Some CRISP-DM activities remain the same such as the activities for business understanding, data preparation, modelling and evaluation. Although some of the associated phases are more appropriately named to align with the philosophy of this framework. The remaining phases in the filtering and mapping section of Figure 4-3 involve the integration of one or more of the activities to address the CRISP-DM shortcomings. For example, the activities of searching for unanticipated trends and anomalies (3), assessing data quality and value (B), detecting broken data (F), detecting bad-quality data (G) and detecting background issues (H) may all be achieved in a data assessment phase. The insufficient data quality monitoring shortcoming, discussed in subsection 4.2.3, is now addressed by creating a data assessment phase. All the phases to carry out the necessary activities for the proactive maintenance of AHUs have now been elicited.

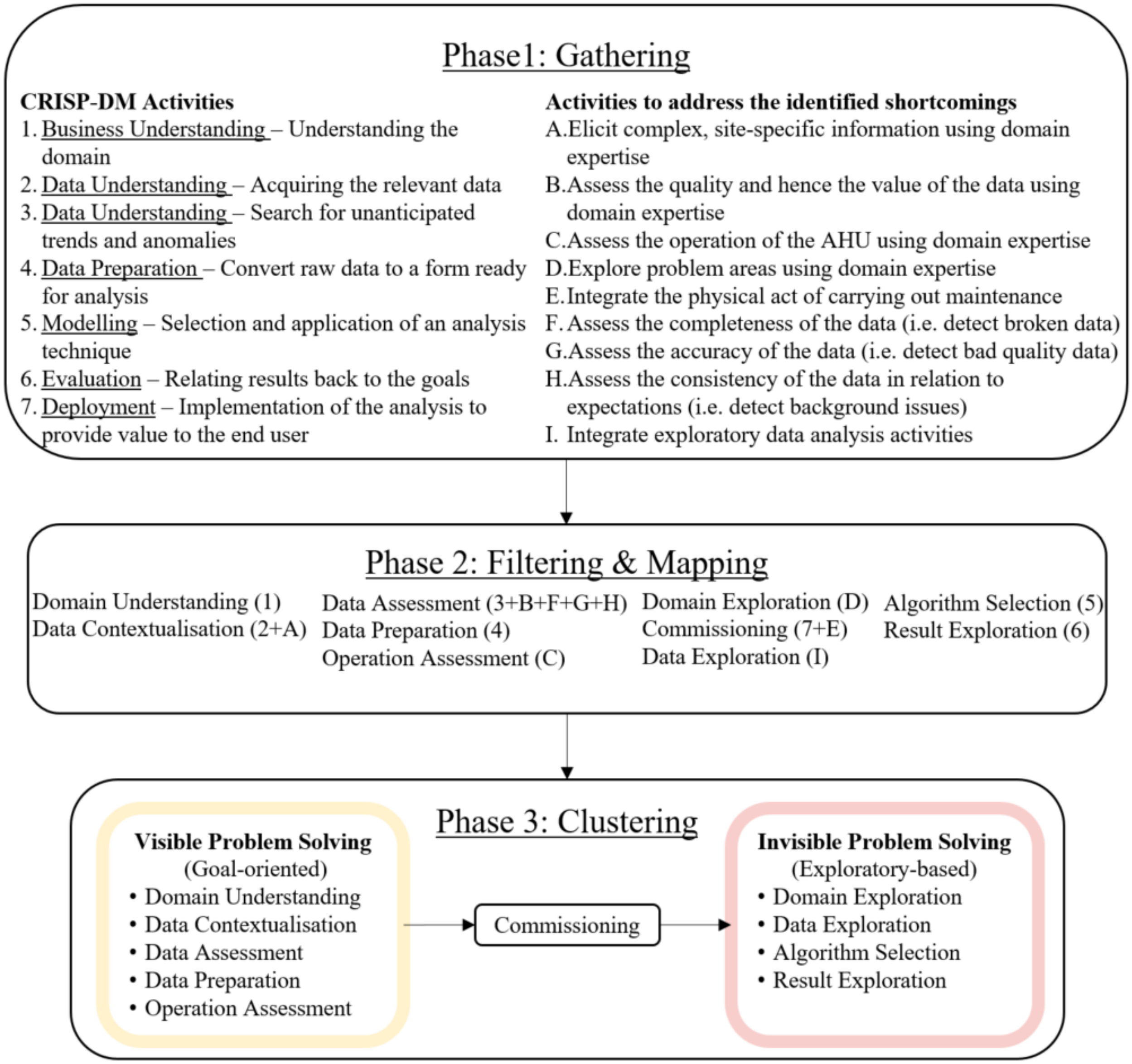


Figure 4-3 Framework Development

4.3.3 Phase 3 - Clustering

Lastly, the newly defined phases are clustered to achieve the goals of the framework. In the discussion of the barriers to proactive maintenance in subsection 2.8, the concept of solving visible or known problems to enable invisible problem-solving was introduced. A tangible means of delivering this concept is shown in the clustering section of Figure 4-3. The phases related to visible, goal-oriented problem-solving are grouped. This then enables known issues to be detected and physically resolved in a commissioning phase.

Subsequently, the activities related to invisible, exploratory-based problem-solving are enabled.

In alignment with the DST map proposed by Martinez-Plumed et al. [87], both goal-oriented and exploratory-based problem-solving are integrated into a cyclical nature. The novel framework developed to house these newly clustered phases is illustrated in Figure 4-4, retaining a similar structure to that of CRISP-DM and the DST map. The framework is named the Industrial Data Analysis Improvement Cycle (IDAIC) [164], as the purpose of the framework is to improve the analysis of industrial data so that industry may transition from reactive to proactive decision-making.

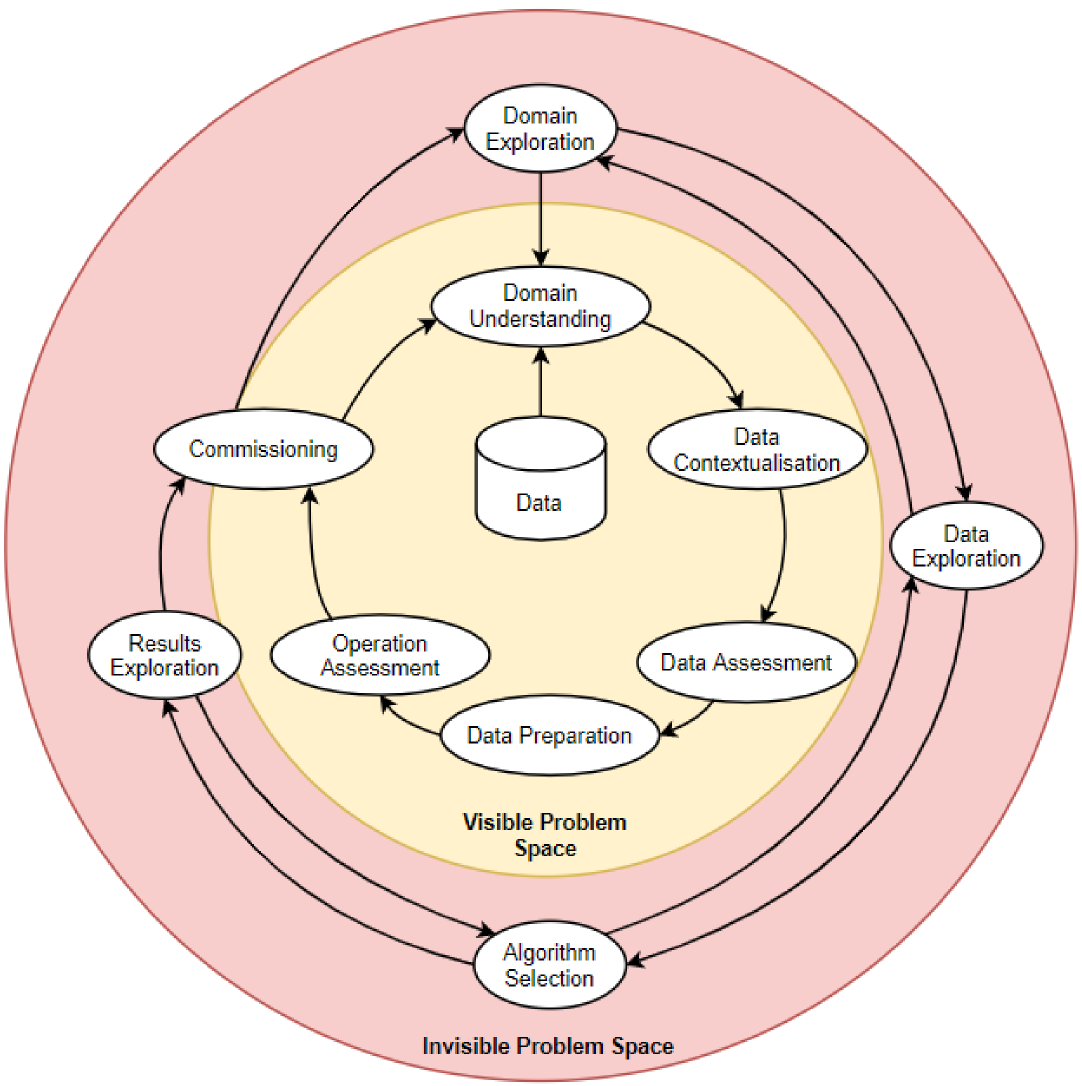


Figure 4-4 Industrial Data Analysis Improvement Cycle

4.4 IDAIC Framework

The lack of direction in terms of the transition to proactive maintenance is addressed in this subsection 4.4 by discussing the phases in the unified IDAIC framework. The phases to solve and avoid visible problems using goal-oriented knowledge-based approaches are discussed in subsections 4.4.1 - 4.4.5. Then, the physical act of maintenance is outlined in the commissioning phase in subsection 4.4.6. Subsequently, the phases to solve and avoid invisible problems using exploratory data-driven approaches are discussed in subsections 4.4.7 - 4.4.10.

4.4.1 Domain Understanding

The first phase in the IDAIC framework aims at collecting the necessary information to obtain a comprehensive understanding of the asset or system under study. Per the Industrial AI paradigm [80], the required inputs may be categorised as people, systems and things. The key people, or the roles that must be fulfilled, are those of a domain expert and a data analyst. An understanding of the domain and an understanding of data analytics are two key skills needed for the successful analysis of industrial data. These people must then integrate with the systems (such as data from heterogeneous sources), and things (such as documentation and equipment specifications). This phase is similar to Business Understanding in CRISP-DM; however, given the engineering context, Domain Understanding is considered to be a more apt name.

4.4.2 Data Contextualisation

The problem understanding that is elicited from the integration of the relevant people, systems and things is then used to contextualize the problem. The relevant data is not only acquired but also evaluated at a high level to ascertain if the data is sufficient. Tasks that would be beneficial to determine the comprehensiveness of the data may include recording the granularity of the data, identifying labels in the data, evaluating the architecture of the data and determining the real-time capabilities of the data. Each of these tasks would aid in the evaluation of the data to determine if the dataset is fit for purpose. This phase is similar to data understanding in CRISP-DM; however, more emphasis is given to determining the problems that the dataset can and cannot help to solve. This is determined through the integration of domain expertise. For example, the availability of key features that are required for expert rule-based approaches offers a preliminary scope to the faults that may be detected. Therefore, the output of this phase is an understanding of the dataset with an emphasis on the identification of the visible problems that may be solved.

4.4.3 Data Assessment

Once the high-level capabilities of the dataset are understood, the data may then be assessed in further detail to evaluate the quality, and hence the value, of the data using domain expertise. A systematic approach is needed to both monitor and manage the "3B's" of industrial data. The area of measuring data uncertainty, or data quality assessment, is not a mature area of research in industrial applications. Cai and Zhu [165] proclaimed that a unified and approved data quality standard had not been agreed upon for big data and that research in the area had just begun. However, many data quality assessment methodologies have been developed in other domains, which have been systematically reviewed by Batini et al. [166]. The authors found accuracy (Acc), completeness (Comp), and consistency (Cons) to be the most frequently monitored dimensions in the literature [166]. Therefore, the assessment of these dimensions is the main focus in this phase, and the associated metrics to assess these dimensions are heavily influenced by domain knowledge. In Figure 4-5, a decision tree is proposed that utilises the common data assessment metrics to monitor the "3B's" and provides recommendations on how to manage them.

Firstly, regarding broken data, if the required data is missing then a digitalisation plan is required to obtain the data. An approach specifically designed for the industrial sector has been proposed by Clancy et al. [167]. This could be quantitatively measured using the percentage availability of key features in the dataset or the percentage availability of each feature for the period of the dataset (Comp). Secondly, regarding bad-quality data, if a particular data point is not in line with the expectation defined using domain knowledge it is inaccurate (Acc). The sensor must then be manually checked and may require recalibration or replacement. Common sensor faults may be identified by using domain expertise to distinguish expected from unexpected values. Finally, regarding background issues with the data, the consistency in the dataset may be monitored. For example, domain expertise may be used to compare similar data values as a means of performing a sanity check or consistency check (Cons). This further contributes to understanding the meaning behind the data, or its usefulness. If the consistency of the data is low, or the data is not per the domain expert's experience, sensors may also need recalibration or replacement. The difference from the accuracy dimension is that inconsistent data may be caused by a component or control fault. Therefore, an operation assessment is needed to determine the source of the anomaly. However before the operation assessment can take place, the data must be appropriately prepared.

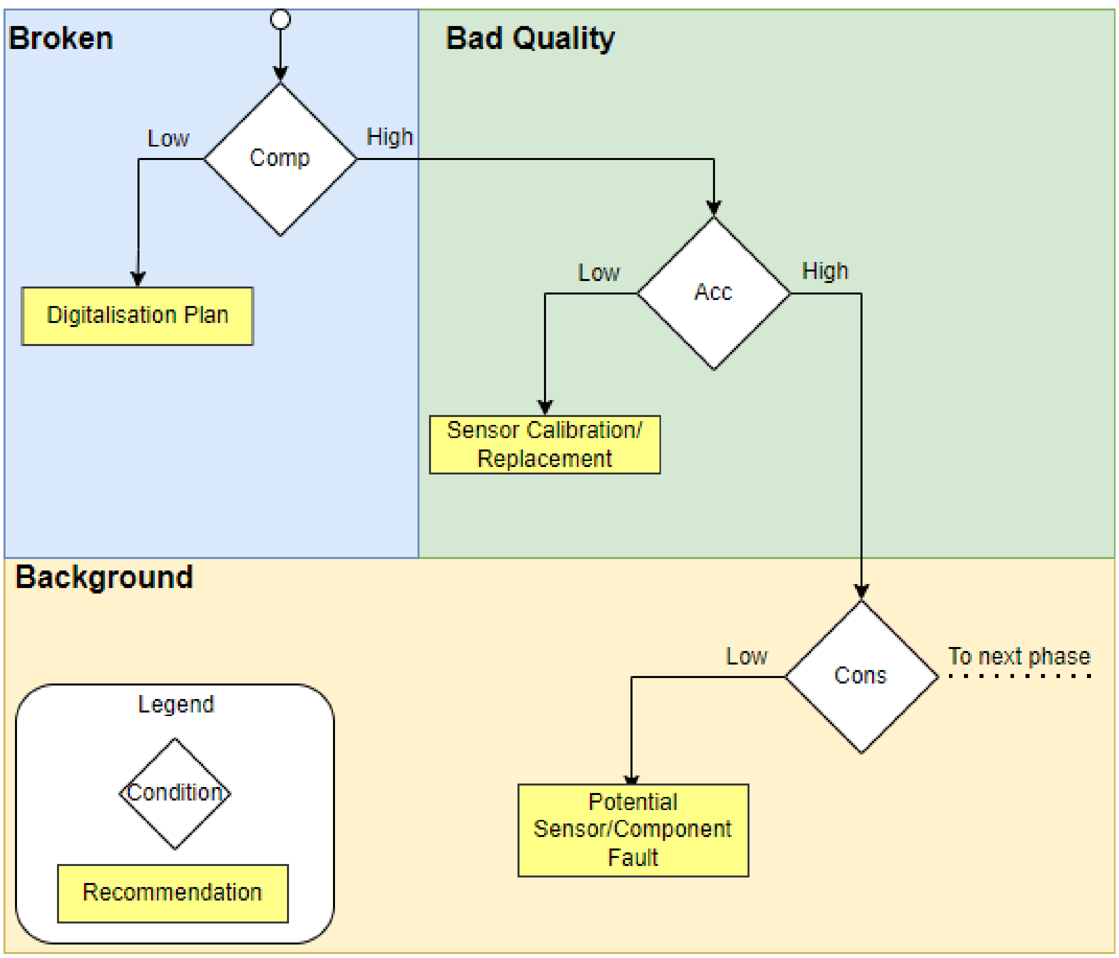


Figure 4-5 Generic Data Assessment Decision Tree

4.4.4 Data Preparation

Data preparation is commonly understood as a laborious phase, consuming approximately 50% or 60% of the total effort in a machine learning project [152]. While a new activity was not added to the data preparation phase during the development of the IDAIC framework, pre-processing tasks are now more informed. For example, issues identified in the data assessment phase uncover nonsensical values that may be removed during the data cleaning process. The integration of domain expertise at this stage has the benefit of increasing the effectiveness of the data preparation, which improves problem-solving performance. Additionally, this proposes a more systematic approach to data pre-processing in building data analytics, which is an area in need of contributions [168].

4.4.5 Operation Assessment

Assessing the operation, or solving visible faults, is aligned with the area of FDD in AHU applications. The visible problems must be solved before the value of data-driven analytics may be realized. The strengths and weaknesses of each FDD classification outlined by Mirnaghi and Haghighat [132] are shown in Table 4-2. Model-based methods, or deep knowledge approaches, are not appropriate for this phase as significant effort is required to develop models that are highly susceptible to error. Data-driven methods are the most promising techniques for highly scalable solutions, but in the visible problem-solving space, the quantity and quality of the data are likely to be insufficient. Therefore, knowledge-based methods, or shallow knowledge approaches, appear to be the best due to their low data requirements. While knowledge-based approaches are limited to the domain expertise available, these methods sufficiently address visible problems in a highly interpretable manner. Knowledge-based methods integrate domain expertise into the data analysis pipeline and ensure the system is operating as expected so that invisible problem-solving is enabled. The operation assessment phase will also aid in the detection of the most problematic components in the system, which may be used to inform the domain exploration phase in the invisible problem space.

Table 4-2 Comparison of strengths and weaknesses of HVAC FDD methods. Reprinted from Mirnaghi and Haghighat [132], *with permission from Elsevier, 2020*

Method	Strength	Weakness
Knowledge-based FDD	• The detailed mathematical model is not needed available • Suitable for the system with a small number of inputs, outputs, and states	• It highly relies on domain expertise while a wide range of faulty and failure cases are beyond engineers' experiences • Not capable of detecting novel failures which are not flagged in the historical data • Cannot work well for complex and large-scale HVAC systems
Model-based FDD	• Works properly when a good HVAC physical model is available	• It is not proper for large-scale and complex HVAC systems • The performance would be limited because of modelling and linearization error
Data-driven based FDD	• Less model development time and cost • No dependency on the model • Easy to retrain • Efficient use of system data • High accuracy • Good performance in works with large-scale and complex systems • minimizing cognitive errors which are beyond the engineers' knowledge	• Need to collect high-quality and sufficient training data for supervised models • High dependency on the quality and the quantity of the collected data

4.4.6 Commissioning

The issues identified in the data assessment and operation assessment must then be actioned upon. In HVAC applications, the manual activity of ensuring systems are operating under design specifications is known as commissioning. In a review of FDD methods for AHUs, Bruton et al. [72] outlined the benefits of FDD-aided commissioning in relation to the four ideal types of commissioning presented in the International Energy Agency's Annex 40 [169]. Depending on the current maintenance strategy of the industrial site, this phase will either involve retro-commissioning or re-commissioning, based on the occurrence of an initial commissioning phase. Then, as the transition to proactive maintenance becomes more advanced, the early detection of faults enables ongoing or monitoring-based commissioning to quickly action maintenance activity. The integration of the commissioning phase is a key adaptation to CRISP-DM as it enables the transition from visible to invisible problem-solving, or reactive to proactive maintenance activities as outlined in Figure 4-6. The issues identified through invisible problem solving are then returned to the commissioning phase so that they may be resolved.

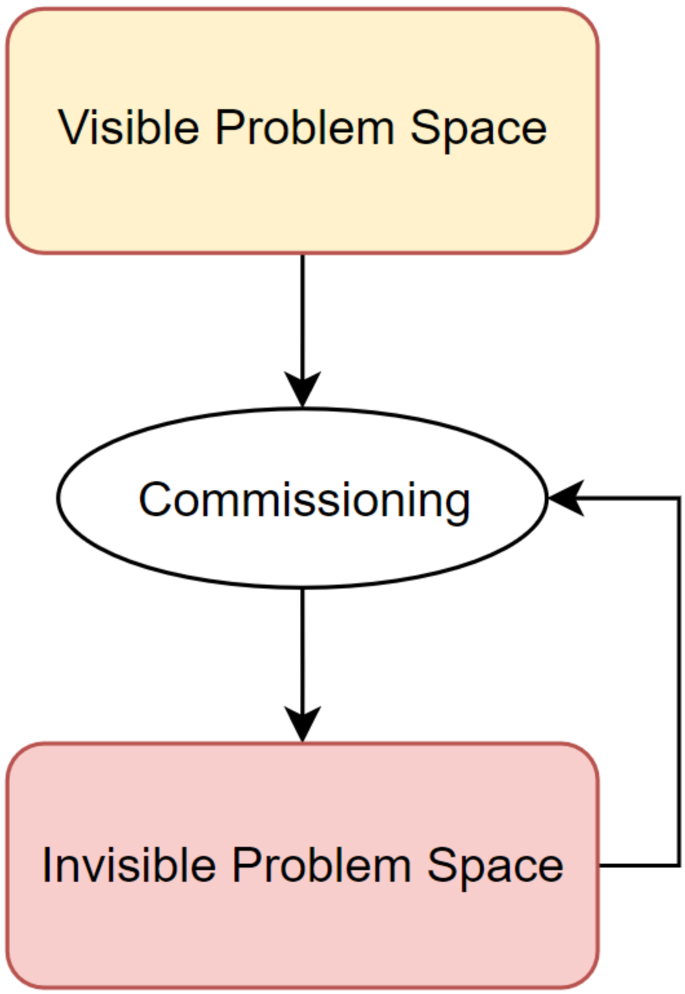


Figure 4-6 Commissioning as the bridge to invisible problem-solving

4.4.7 Domain Exploration

Once the system is re-commissioned or retro-commissioned, the invisible problem-solving activity is enabled. The outer ring in Figure 4-4 aims to explore the data to detect early signs of degradation that would inform decision-makers on the remaining useful life of components to plan maintenance activities accordingly. The first phase is to determine if the current problem understanding is sufficient. To do this, an exploration of the domain is undertaken to establish if the system is operating stably and efficiently. While the operation assessment aims at detecting and diagnosing visible problems, the domain exploration aims at assessing the efficiency of the system and identifying the areas where data-driven analysis may provide value. For example, the operation assessment may identify a passing valve that is then replaced in the commissioning phase. However, this valve may have reached the end of its useful life prematurely, resulting in costly replacement and insufficient knowledge as to the cause of the rapid degradation of this valve. In this scenario, the data-driven analysis may provide insight into the degradation pattern of the new valve, creating valuable information for the facilities management team that enables smarter, proactive decision-making.

4.4.8 Data Exploration

In the following phase, the proposed concepts, or areas for further exploration, are assessed using the data. The viability of a data-driven approach is justified based on the ability of the dataset to address the invisible problem. Common exploratory data analysis and ML techniques are tested in this phase to uncover hidden relationships or unexpected patterns or anomalies in the data that may substantiate the signs of an invisible problem. Continuing the example of analysing the valve degradation, the relationship between key features may be presented visually by the data analyst and interpreted by the domain expert. Depending on the form or strength of the relationship, a regression model may be fitted to the data or an anomaly detection algorithm may be more suitable. While the specific algorithm is selected in the next phase, this phase enables a dialogue between the data analyst and domain expert to explore the invisible problem-solving capabilities for the AHU.

4.4.9 Algorithm Selection

The most appropriate algorithm or technique is then selected in the penultimate phase. At the high level, ML algorithms may be classified as supervised, unsupervised, semi-supervised or reinforcement. Based on the findings in the data exploration phase, an appropriate algorithm is selected that accounts for domain-specific auxiliary information that may not be captured in the data.

4.4.10 Results Exploration

Lastly, the meaning of the results is explored. The output of the data-driven analysis is critically appraised before an action item is raised for the commissioning phase. In this manner, the visible and invisible problem-solving tasks are integrated into a single framework facilitating continuous improvement.

4.5 Overview of the IDAIC Framework contributions

The main novel contributions of this framework to address the CRISP-DM shortcomings are presented in Table 4-3. Firstly, the knowledge gap is addressed through the development of IDAIC which integrates domain expertise into the data analysis process to overcome the barrier of a lack of domain-guided analysis. A framework that captures the occurrence of hidden faults will enable the management of a facility to get greater insight into daily operations. As a result of improved visibility to management, facility technicians will be motivated to promptly resolve issues given that they may be credited for doing so. Secondly, the domain understanding, domain exploration, and operation assessment phases provide dedicated phases to elicit the available domain expertise to address the lack of knowledge of the domain from the data analyst's perspective. Thirdly, the lack of integration of the data analysis with the physical process of maintaining the AHU is addressed through the commissioning phase. Fourthly, insufficient data quality monitoring is addressed in the data assessment phase. A means of monitoring each of the three main data issues (3B's) is outlined in Table 4-4. This provides a quantifiable measure of the value of the available data. Finally, the lack of exploratory-based analysis is addressed by the invisible problem-solving track. This enables problem areas to be uncovered and investigated using data-driven approaches.

Table 4-3 Novel Contributions for each shortcoming of CRISP-DM for proactive maintenance of AHUs

Shortcoming	Novel Contribution	Purpose	Explanation
Lack of domain knowledge	Domain Understanding	To acquire the inputs for reactive analysis	Identifies the key personnel and data sources
	Domain Exploration	To acquire the inputs for proactive analysis	Integrates the undocumented operational history of the AHU known to onsite personnel
	Operation Assessment	To detect the occurrence of visible faults	Integrates knowledge-based FDD approaches to identify the ground state
Lack of integration with existing strategies	Commissioning	To resolve faults	Integrates the manual process of resolving faults in AHUs with the analysis
Insufficient data quality monitoring	Data Assessment	To detect data issues	Identifies issues with the data in relation to broken, bad-quality and background issues
Lack of exploratory-based analysis	Invisible Problem-Solving Track	To detect the occurrence of invisible faults	Visible track for goal-oriented analysis and invisible track for exploratory-based analysis

Table 4-4 Overview of the data assessment process

Data Issue	Dimension	Metric	Explanation
Broken	Completeness	Availability	Number of problems that can/cannot be solved with the data available using an expert rule set
Bad-quality	Accuracy	Syntactic Accuracy	Single feature analysis to detect outliers and stuck values
Background	Consistency	Semantic Consistency	Multiple feature analysis to detect inconsistencies in the data (e.g. drift, bias, and precision degradation)

4.6 Discussion

A novel framework has been developed to transition to proactive maintenance on AHUs by advancing the analysis of industrial data to reduce energy consumption. Traditionally, data analysis has relied on the data analyst's ability to understand the domain under investigation. However, to achieve more accurate and insightful results in the context of maintaining AHUs, a novel adaptation of CRISP-DM was needed. The Industrial Data Analysis Improvement Cycle (IDAIC) depicted in Figure 4-4 is an integration-type adaptation that incorporates data assessment techniques, expert systems, commissioning, and exploratory data analysis to bridge the knowledge gap. The framework delivers both short- and long-term value by identifying faults through knowledge-based approaches, while also evaluating data quality and monitoring its progress over time.

The IDAIC framework begins with understanding-centric phases, whereby the domain and the data are assessed to determine the value that the problem-solving activities may realise. Data assessment is then performed to determine the quality of the dataset or the fitness of the dataset for AI & ML applications. Data points that are deemed to be of concern are marked for further analysis in the operation assessment phase, with appropriate pre-processing activities performed in advance. The operation of the AHU is assessed using knowledge-based FDD techniques, such as expert rules, to identify visible problems in the system. While problems that are identified may be caused by those bad-quality data points marked from the data assessment phase, it may incentivize industrial practitioners to recalibrate or replace faulty sensors when highly interpretable expert rules cannot be applied. While the IDAIC framework aims to enable industry to perform advanced analysis, the overall goal is energy consumption reduction. Therefore, the analysis must inform the physical changes that must be made to the system to improve its energy performance. Commissioning may be defined as "a quality-oriented process for achieving, verifying and documenting whether the performance of a building's systems and assemblies meet defined objectives and criteria" [169]. While sensor faults are resolved in commissioning activities, the data quality demands of AI and ML solutions require rigorous data assessment and maintenance of the sensor. Hence, the commissioning phase within the IDAIC framework serves a dual purpose. It not only performs the necessary maintenance activities to achieve energy-efficient operation but also aims at performing the necessary maintenance activities to collect the highest quality of data possible.

At this point in the IDAIC framework, domain knowledge has been sufficiently integrated to solve visible problems and the AHU is operating per a domain expert's expectations, with good-quality data continuously being collected. Industry may then switch from a reactive maintenance approach to a proactive maintenance approach with data-driven techniques. However, while domain expertise informed the inner goal-oriented circle, the outer circle is data-oriented and exploratory in nature. Similar to the DST map in Figure 2-11 proposed by Martinez-Plumed et al. [87], invisible problem-solving and avoidance begin with a goal, or domain, exploration. Strong collaboration is required between data analysts and domain experts in this phase to elicit potential areas of interest to apply data-driven algorithms. For example, the cause of a recurring fault or failure may be unknown and an unsupervised ML algorithm may provide insight through pattern recognition or clustering techniques. Similarly, ML may be used to anticipate a particular fault that is deemed critical by the domain expert, providing an estimate of the remaining useful life in a predictive maintenance strategy. Once a goal has been selected, data exploration takes place to uncover if all relevant information is available and of sufficient quality to perform the analysis. If this is not the case, a return to the domain exploration phase is needed to either update the goal based on the information available or else return to the domain understanding phase to obtain this information. Data exploration is a necessary intermediary phase to determine if the aspirations of the defined goal are attainable. If the data-oriented goal is achievable, an appropriate algorithm must then be selected. The challenge of algorithm selection and result evaluation is commonly stated in the literature as shown in Table 2-15. However, contributions to the field this year have begun to offer guidance in this regard as in Figure 2-15. By solving the visible problems and determining if the data-oriented goal is achievable in the domain exploration phase, much of the uncertainty in the data-driven analysis will be removed, enabling a clearer decision to be made on algorithm selection. The algorithms' results are then evaluated in the final phase of the IDAIC framework, whereby the results are used to inform decision-makers to make energy reduction decisions.

Regarding the practical use of IDAIC, there are many arrows outlined in Figure 4-4. For example, the invisible problem-solving space is exploratory in nature, therefore, backward arrows are included to denote the back-and-forth nature of this process. While this is also the case in the visible problem-solving space, the latter phases of the IDAIC framework are less prescriptive and more interoperation between phases may be necessary. Furthermore, not every possible transition is shown in Figure 4-4. Similarly to the DST map proposed by Martinez-Plumed et al. [87], not every phase must be completed for successful industrial data analysis. The IDAIC framework merely provides a structured and flexible framework for early adopters of digitalisation to manage the transition to the data-driven analysis of industrial data. It is not designed to eradicate every challenge faced in industrial data analysis through rigid implementation. For example, in systems that are well maintained, visible problems may not occur rendering the operation assessment phase redundant. It is therefore developed for guidance rather than regulation.

4.7 Conclusion of IDAIC Development

A conceptual framework has been developed to aid the transition from reactive to proactive maintenance approaches to enable energy reduction. An integration-type adaptation of CRISP-DM is developed to incorporate domain expertise into the analysis. The novel integration of engineering knowledge and a data science process model has led to a structured cycle whereby knowledge-based approaches solve and avoid visible problems, and data-driven approaches solve and avoid invisible problems. Therefore, the novel IDAIC framework addresses the lack of direction to transition from reactive to proactive maintenance approaches. The second novel contribution in this chapter relates to the integration of domain knowledge, such as through knowledge-based FDD in the operation assessment phase. This facilitates retrospective analysis of the AHU to isolate visible problems that must be solved before the value of data-driven analysis may be realised in invisible problem-solving. Thirdly, commissioning is added to integrate the physical process of carrying out maintenance. Fourthly, a data assessment phase is added to manage the "3B's" data issues. Specific contributions include a completeness review to understand the faults that may be detected with the data, an accuracy review to identify sensors in need of calibration and replacement, and a consistency review to detect data anomalies indicative of subtle sensor faults. Lastly, data-driven analysis is enabled through the invisible problem-solving space.

While the proposed IDAIC framework has been discussed at a high level, a practical implementation must follow. As noted by Wirth and Hipp [88], "If the early adopters fail with their data mining projects, they will not blame their own incompetence in using data mining properly but assert that data mining does not work". The IDAIC framework must now be applied to an air handling unit in an industrial facility to test the applicability of the framework in the real world.

5 CHAPTER 5 - FRAMEWORK IMPLEMENTATION

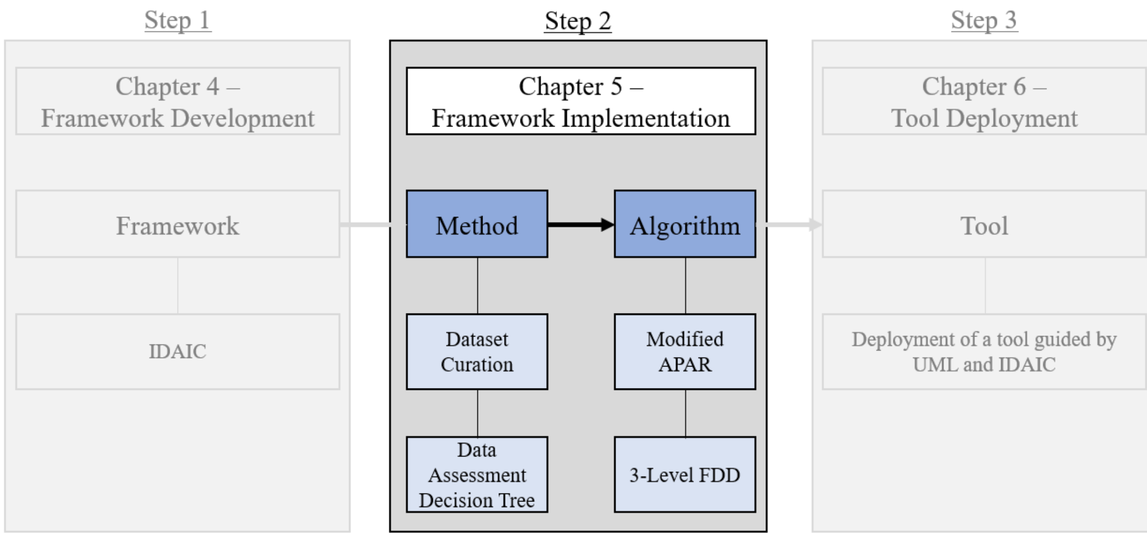


Figure 5-1 Step 2 of the research methodology

5.1 Introduction

The objective of this chapter is to document the implementation of the IDAIC framework at a large manufacturing company. This would address the second research question, "*How can raw AHU data from the industrial setting be systematically curated into a dataset that enables the development of practical and reproducible data-driven analysis?"*. Additionally, the third research question will also be addressed, *"How can the operation of a real-world AHU in an industrial facility be assessed to identify the sensor, component, and control faults that must be addressed prior to the application of data-driven approaches for the purpose of proactively maintaining the AHU?"*

The aim is to show the usefulness and practicality of the IDAIC framework which has multiple novel contributions that address the shortcomings of CRISP-DM for proactive maintenance of AHUs. However, specific emphasis is given to demonstrating its effectiveness in bridging the data and operational gaps to address the second and third research questions. The implementation begins with the goal-oriented visible problem-solving process that includes dataset curation, data assessment, and the identification of those faults that are known to domain experts. This is discussed in subsection 5.2. The resolution of these faults then enables the application of data-driven approaches to identify those faults that are not as easily identifiable (invisible problems) which is discussed in subsection 5.3. The results and impact of the analysis are then discussed in subsection 5.4, with conclusions stated in subsection 5.5. However, firstly, in subsection 5.1.1, a concise background of the case study site is presented.

5.1.1 Background

The case study site selected to implement this framework is part of a multi-national organisation and is located in county Cork, Ireland. The site was built in 1997 for manufacturing medical devices. Multiple AHUs, chillers, and boilers have been commissioned on this site to provide adequately ventilated and thermally comfortable working environments. One of the largest air volume AHUs that serves a production floor space in the facility is selected for this case study. This particular AHU (Figure 5-2) is over twenty years old, operates 24/7, and is one of four similar AHUs which contribute to its selection as an appropriate case study with potential for replication. This AHU is a return air unit and serves an open production floor from thermal comfort and an adequate supply of fresh air perspectives only. The data is collected at 15-min intervals which is representative of the granularity of real-world systems, given often local Building Management System (BMS) based data storage constraints. Multiple air temperature sensors are located throughout the AHU as detailed in Figure 5-2 and are utilised by the BMS for control purposes. As a result, the data was originally collected for this specific purpose and not intended for advanced data analysis. Consequently, the control strategy only utilises instantaneous data values, leaving a large quantity of historical data untapped from a data analysis perspective. With the development of the IDAIC framework, there is now a significant opportunity to leverage this historical data to gain greater insight into the AHU’s operation, thereby enabling actions to be taken to reduce its associated energy consumption.

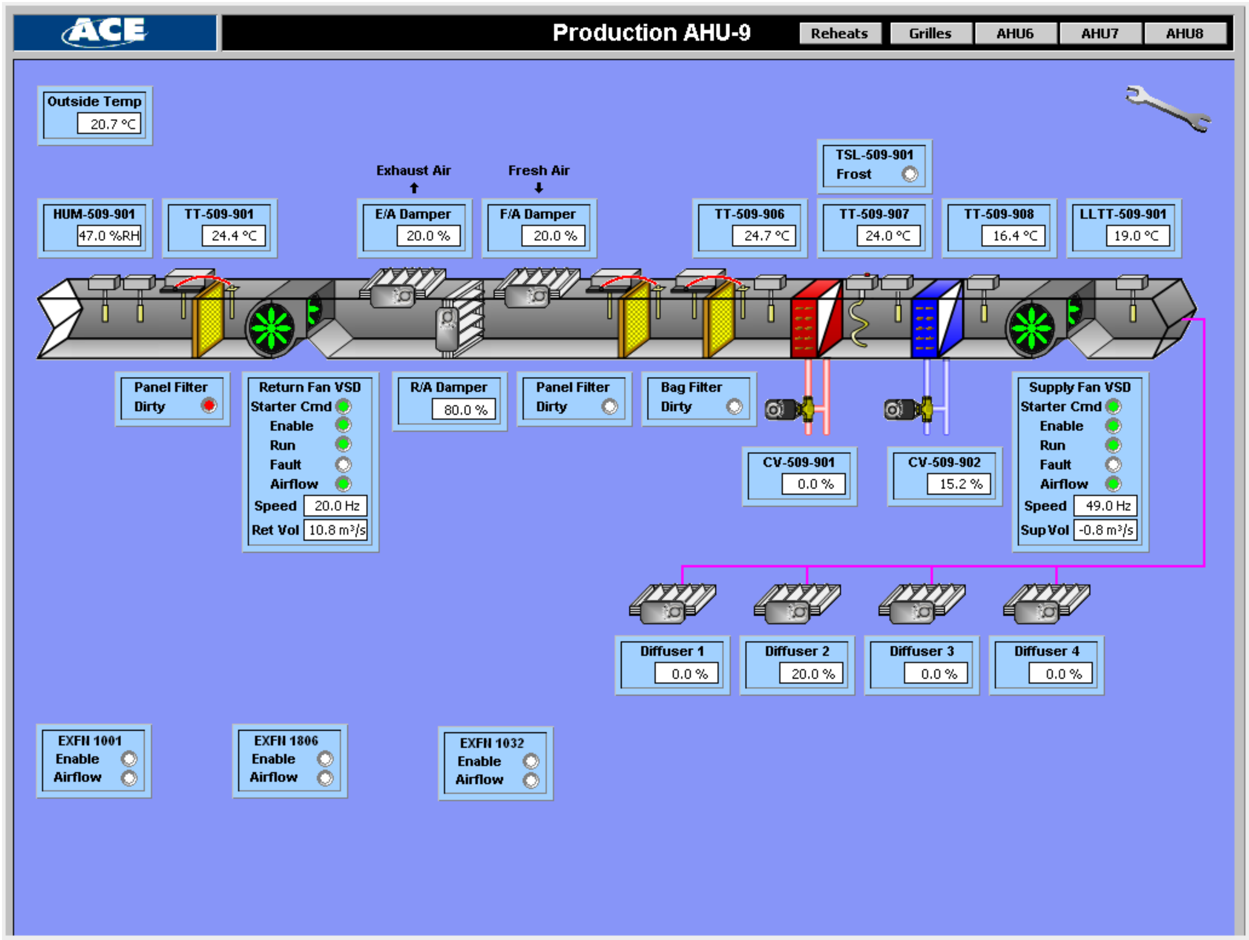


Figure 5-2 Screenshot of BMS screen of the case study AHU

5.1.2 Research Objectives

There are two main objectives in this chapter. Firstly, to curate a dataset from raw BMS data that is representative of the challenges in the industrial setting, and assess it, to address the data gap. Secondly, implement the IDAIC framework on this AHU to detect both component and control faults to address the operational gap.

5.2 Implementation of IDAIC for Visible Problem-Solving

5.2.1 Domain Understanding

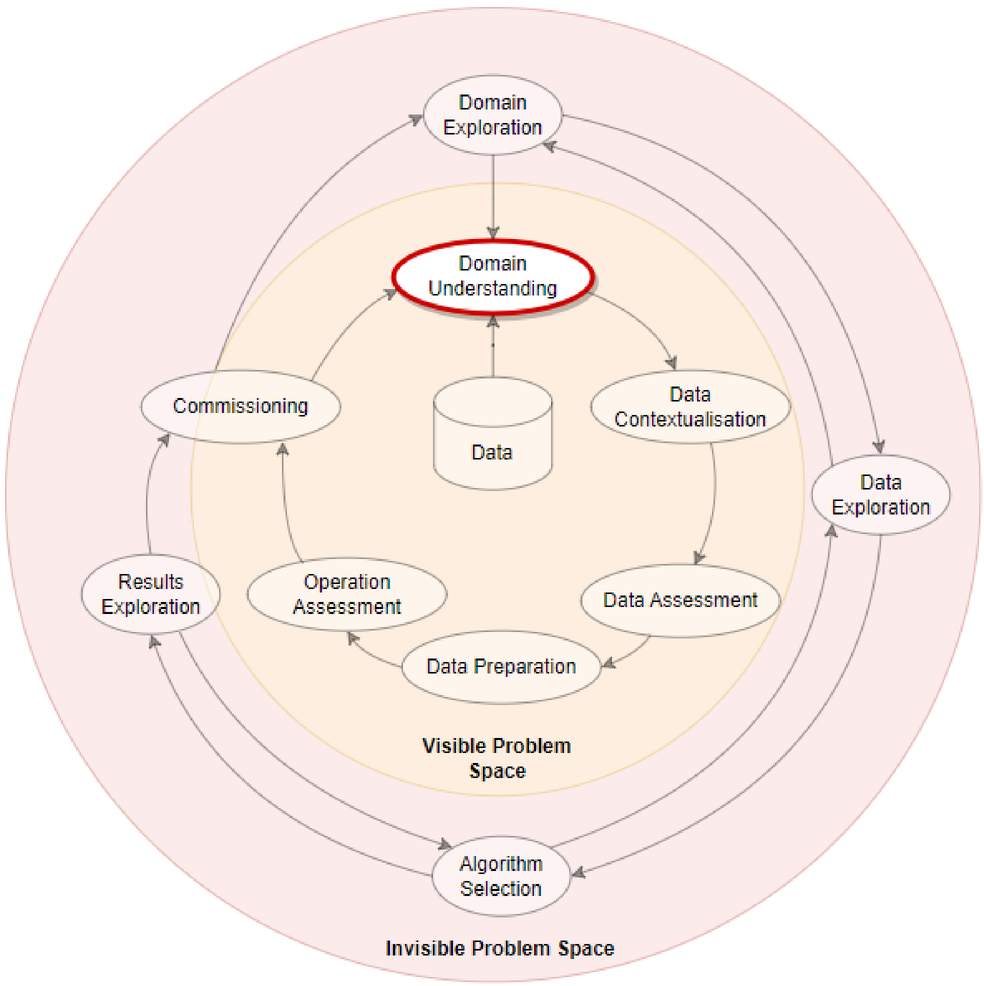


Figure 5-3 Domain Understanding Phase of the IDAIC Framework

The first phase, domain understanding, aims to integrate the necessary people, systems and things. This refers to the incorporation of the key stakeholders, the processes or software, and the relevant physical objects respectively. Firstly, the necessary roles to be filled include those of a domain expert and a data analyst. The challenge lies in acquiring site-specific knowledge, which is crucial for meaningful analysis and requires input from onsite personnel. To obtain this site-specific knowledge, stakeholder meetings were conducted with the onsite facilities. These meetings occurred weekly so that regular communication was established, leading to strong integration of onsite knowledge in the analysis. The progression through the phases of the IDAIC framework provided guidance and structure for these meetings, eliciting key information from the facilities team as outlined in upcoming phases. Secondly, the main system that stores the data is the BMS. The data that is relevant to this AHU is presented in Table 5-1 and Table 5-2, with their physical location depicted in Figure 5-4. Thirdly, the "thing" generally refers to document review, which in the case of this site is provided by the functional design specification (FDS). The FDS outlines the modes of operation of the AHU and contains information regarding the location of sensors in the unit. The production floor area that this AHU serves is split into two zones, zone 1 and zone 2, each with a different load profile and accompanying reheat coils. The presence of reheat coils adds another layer of complexity, therefore the analysis of the visible problem space is confined to the sensors and components within the AHU as shown in Figure 5-4. The remaining sensors and components in Table 5-2 are then included in the invisible problem-solving task in subsection 5.3.

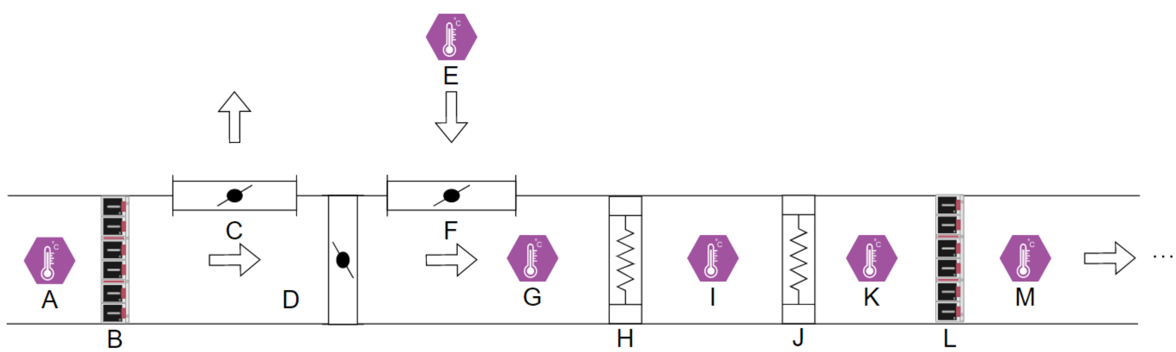


Figure 5-4 Schematic of the AHU

Table 5-1 Data point names, descriptions and expected region of values within the main unit

** Excluded from the analysis as OaDmprPos contains the necessary signal information*

Location in Figure 5-4	Point Name	Description	Expected Region
A	RaTemp	Return air temperature	[15, 25] °C
B	RaFanPower	Return air fan power	[0, 30] kWh
C	EaDmprPos*	Exhaust air damper control signal	N/A
D	RaDmprPos*	Return air damper control signal	N/A
E	OaTemp	Outside air temperature sensor	[-10, 25] °C
F	OaDmprPos	Outside air damper control signal	[20, 100] %
G	MaTemp	Mixed air temperature	[5, 25] °C
H	HWVlvPos	Hot water valve control signal	[0, 100] %
I	HCALTemp	Air leaving hot water coil temperature	[5, 25] °C
J	ChWVlvPos	Chilled water valve control signal	[0, 100] %
K	CCALTemp	Air leaving chilled water coil temperature	[5, 25] °C
L	DaFanPower	Supply air fan power	[0, 30] kWh
M	DaTemp	Supply air temperature	[10, 25] °C

Table 5-2 Data point names, descriptions and expected region of values beyond the main unit

Location	Point Name	Description	Expected Region
Nearby Weather Station	OaTemp_WS	Outside air temperature of nearby weather station	[-10, 25] °C
Zone 1 Supply Duct	ReHeatVlvPos_1	Reheat hot water valve control signal for zone 1	[0, 100] %
Zone 2 Supply Duct	ReHeatVlvPos_2	Reheat hot water valve control signal for zone 2	[0, 100] %
Zone 1 Supply Duct	ZoneDaTemp_1	Zone supply air temperature for zone 1	[15, 25] °C
Zone 2 Supply Duct	ZoneDaTemp_2	Zone supply air temperature for zone 2	[15, 25] °C
Zone 1	ZoneTemp_1	Average zone 1 air temperature	[15, 25] °C
Zone 2	ZoneTemp_2	Average zone 2 air temperature	[15, 25] °C

Furthermore, the faults of interest for the visible problem-solving space must be defined in this phase. The most commonly referred to faults in systematic reviews of AHU FDD [19], [139], [170] are categorised as either visible or invisible faults in Table 5-3. The classification is based on the nature of the fault. Faults that have a strong distinctive characteristic that may be easily detected are classified as visible, whereas subtle degradation-based faults are classified as invisible. Sensor faults are closely related to data issues and are therefore addressed in the data assessment phase. Given that invisible faults are often a precursor to visible faults, they are closely linked and the classification may change as the analysis progresses.

Table 5-3 Faults of Interest

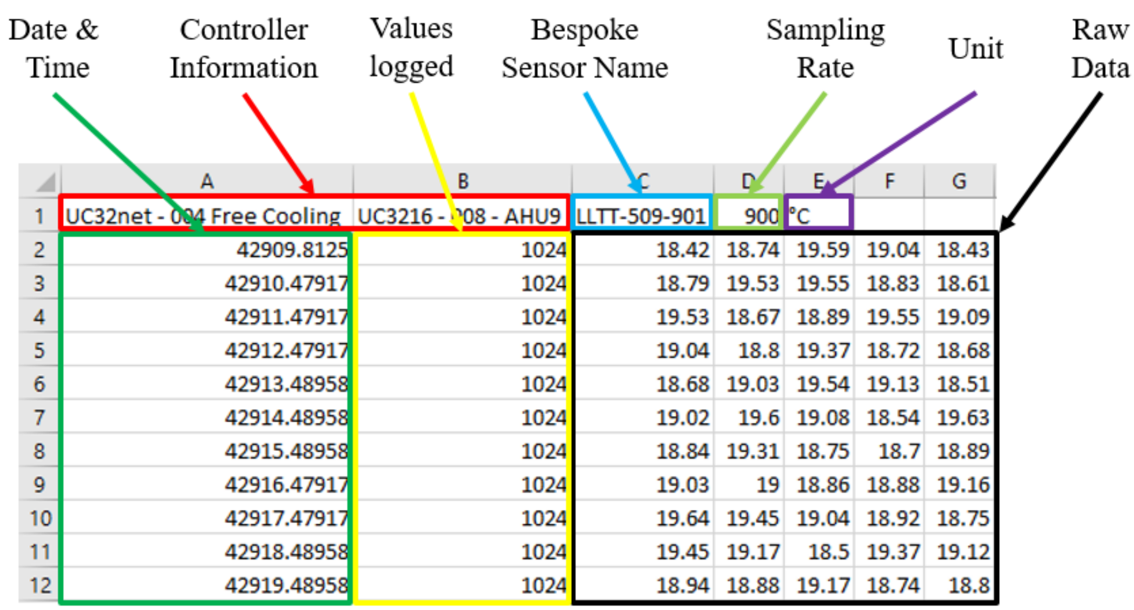
Category	Device	Fault	Visible	Invisible	Rule
Equipment	Fan	Overall failure	X		N/A
		Decreased motor efficiency		X	N/A
	Heating Coil	Fouling		X	3,4
		Water pump failure	X		1,3,4
	Cooling Coil	Fouling		X	12,13,14, 17,19,20
		Water pump failure	X		11,12,13, 14,16,17, 19,20
Actuator	Dampers	Stuck in a faulty position	X		2,8,10,18
		Air leakage		X	2,8,10,18
	Cooling Coil Valve	Stuck, broken or wrong operation	X		1,3,4,7,11,12,13,14,16,17,19 20
		Leakage		X	1,3,4,7
	Heating Coil Valve	Stuck, broken or wrong operation	X		1,3,4,7,8,11,12,13,14,16,17, 19,20
		Leakage		X	7,8,11,12,13,14,16, 17,19,20

Category	Device	Fault	Visible	Invisible	Rule
Sensor	Temperature and Flowrate	Analysed in the Data Assessment phase			
Controller	Sequence of heating and cooling valves	Unstable response	X		28
	Zone temperature	Unstable response	X		25

5.2.2 Data Contextualisation

Figure 5-5 Data Contextualisation phase of the IDAIC framework

In the second phase, the data is contextualised by acquiring the relevant data and then assessing its adequacy concerning the number of expert rules that may be applied. The data was obtained in two main ways. Firstly, the Cylon Unitron BMS [171] archives a log of the data on the physical PC running Windows XP Professional. During this archival process, data from the previous 10 days are appended to CSV files located in an archive folder containing 696 files, one per monitored point on the entire BMS control system. Each logged data point has its own CSV file in the folder that is uniquely named, without any contextual information. The only metadata available is contained on the first row of the CSV file. An example of one such CSV file is shown in Figure 5-6. In cells A1 and B1, information regarding the controller for this sensor is logged. From this information, it may be deduced that this data point relates to AHU9 as in cell B1. The bespoke name of the sensor is contained in cell C1. In the FDS, this point refers to AHU Duct Mounted Supply Air Low-Level Temperature Transmitter. Therefore, the beginning of the identifier "LLTT" stands for "Low-Level Temperature Transmitter". Cell D1 then records the sampling rate of 900s or 15 minutes. Finally, cell E1 contains the unit of the sensor.



	A	B	C	D	E	F	G
1	UC32net - 0[illegible]4 Free Cooling	UC3216 - [illegible]08 - AHU9	LLTT-509-901	900	°C		
2	42909.8125	1024	18.42	18.74	19.59	19.04	18.43
3	42910.47917	1024	18.79	19.53	19.55	18.83	18.61
4	42911.47917	1024	19.53	18.67	18.89	19.55	19.09
5	42912.47917	1024	19.04	18.8	19.37	18.72	18.68
6	42913.48958	1024	18.68	19.03	19.54	19.13	18.51
7	42914.48958	1024	19.02	19.6	19.08	18.54	19.63
8	42915.48958	1024	18.84	19.31	18.75	18.7	18.89
9	42916.47917	1024	19.03	19	18.86	18.88	19.16
10	42917.47917	1024	19.64	19.45	19.04	18.92	18.75
11	42918.48958	1024	19.45	19.17	18.5	19.37	19.12
12	42919.48958	1024	18.94	18.88	19.17	18.74	18.8

Figure 5-6 Example of a CSV file located in the archive folder of the BMS

The remaining rows of the CSV file contain the daily log of data for the previous 10 days, with each point representing one 15-min value. The timestamp information is contained in column A. The entries are stored such that the whole number represents the date and the decimal represents the time. In cell A2, the date represents a value that corresponds to 42,909 days after January 1, 1900, along with a fraction of 0.8125 representing a portion of the day. Therefore, this entry refers to the 25th of June 2017 at 19:30:00. This is the time at which the first value in cell C2 was recorded. Cell B2 then contains the number of values logged on that date. Given the 15-min sampling rate, 1024 values equate to 10 days and 16 hours of data. In row 3 the process is repeated for data recorded the following day.

There is also another means of obtaining this data. As of 2019, a “JACE” (Java Application Control Engine) controller [172] was deployed on the BMS. This is a hardware device produced by Tridium Inc. [172], a company that specialises in building automation and control systems. This enabled a third-party company to access the data in real-time (every 15 minutes). The third-party company were granted read-only privileges for this data as part of this research. It also broadened the scope of the analysis to include data sources other than the BMS, such as fan energy meter data that is located on the site’s proprietary OSI PI historian [173].

While the data comes from the same physical sensors, there are some differences between the Cylon and Tridium data. One example is the time of the sample. While both systems should contain data that is timestamped on the quarter hour, there is potential for some discrepancies to exist such as a 30-second lag between recorded values. Furthermore, there is a difference in the naming convention between Cylon (cell C1 in Figure 5-6) and the bespoke name applied by the third-party vendor. Standardising the data tagging is therefore key to successfully integrating and leveraging the data from both of these sources. This is a crucial component of performing reproducible and practical analysis. Efforts are currently underway to standardise semantic tags for building data in the proposed ASHRAE Standard 223P [174], which is a collaboration between Project Haystack [175] and Brick Initiative [176]. The Department of Energy in the US has allocated $1,300,000 for the financial year 2023 with the project term set from October 1st 2023 to September 30th 2025 [177]. Research centres that are active in the literature with highly cited publications are partners in this project, including the National Renewable Energy Laboratory – Golden, CO [178], Lawrence Berkeley National Laboratory – Berkeley, CA [179], and Pacific Northwest National Laboratory – Portland OR [180]. This would provide the necessary guidance for large-scale reproducible analysis. However for this case study, a single AHU is under investigation, therefore aligning the sensor names to standardised naming conventions such as Project Haystack was deemed sufficient.

To contextualise and align the two sources, a mapping file using Project Haystack was created. An example is shown in Figure 5-7 for the file in Figure 5-6. The asset, AHU9, is mapped to the log file name, sensor identifier and point identifier used by the third-party company. This is then standardised per the Project Haystack naming convention as outlined in the State of Utah Haystack Tagging Reference Model Example [181]. The remaining data points relating to this AHU are similarly contextualised as presented in Table 5-1.

	A	B	C	D	E	F	G
1	**Asset**	**No.**	**FDS Description**	**Log File Name**	**Log File Identifier**	**Third Party Identifier**	**Project Haystack Name**
2	AHU	09	Duct Mounted Supply Air Low Level Temperature Transmitter	D0040801	LLTT-509-901	AHU-09, Unit Supply Air Temperature, ºC, Production, Plantroom	DaTemp

Figure 5-7 Example of a mapping file entry

An overview of the Data Contextualisation phase, referred to as dataset evaluation, is presented in Table 5-4. The key high-level characteristics of the data are recorded to determine its adequacy. For example, the experimental studies identified in the literature generally utilise 1-min granularity data along with data labels. This provides rich information given that 60 data points are recorded in one hour, enabling many different forms of analysis to take place such as supervised, unsupervised and semi-supervised ML. As the data in the industrial site is unlabelled and much less granular at 15-min intervals, only 4 data points are obtained in one hour with an unknown operating state. Therefore, the 15-min intervals and lack of labels are inadequate in their current form for some analyses such as supervised ML. However, given that the visible problem-solving space utilises shallow knowledge rules with low data requirements, of which 15-min data has been shown to successfully identify faults by Bruton et al. [182], the dataset is deemed adequate. Furthermore, as the data spans over 5 years and contains 7 of the 11 key features needed to apply the APAR ruleset as in Table 5-5, it may be considered adequate for further analysis, although deficient in parts. While only 7 of the 11 features would appear to be low, appropriate assumptions may be made through the integration of site-specific knowledge. Firstly, the occupancy status is not required as the zone is always occupied due to 24/7 operation. Secondly, the AHU is controlled to the zone air temperature set point which may be assumed constant. While the final two features, the return air and outdoor air relative humidity, may be ignored as this unit is only designed for temperature control. This means that all the key features are available for the application of APAR.

The selection of APAR for visible problem-solving is discussed further in the upcoming operation assessment phase.

Table 5-4 Dataset Evaluation

Data Granularity	15-min intervals
Data Labels	None
Data Architecture	Cylon BMS archived log files & 3rd party cloud-based database accessed via the Tridium JACE box
Real-time capabilities	Non-real-time (Cylon log files) and near real-time (Tridium JACE box)
Data Adequacy	Adequate for shallow knowledge rules

Table 5-5 Key features for APAR application identified by Schein et al [183]

Requirements for APAR	Status
Occupancy Status	Not required as the production floor is in a 24-hour operation.
Supply air temperature set point	The AHU is controlled to an average zone air temperature set point which may be obtained from the BMS but is not recorded.
Supply air temperature	Available
Return air temperature	Available
Mixed air temperature	Available
Outdoor air temperature	Available
Cooling coil valve control signal	Available
Heating coil valve control signal	Available
Mixing box dampers control signal	Available
Return air relative humidity	For enthalpy-based economizers only
Outdoor air relative humidity	For enthalpy-based economizers only

Given that the data is stored in heterogeneous sources and the log files are not in a form ready for analysis, some initial pre-processing or data preparation was required. While Data Assessment is the next phase in the IDAIC framework, the CRISP-DM properties are retained such that tasks may be performed in a different order and certain actions repeated [88].

5.2.2.1 Initial Data Preparation

Given the objective of this research is to create practical and reproducible analysis, a transparent means of preparing, analysing and visualising the data is needed. As outlined in reviews of the AHU FDD literature, TRNSYS, SIMBAD, EnergyPlus, and OpenModelica are popular software selections [105]. While these are powerful tools for energy system simulation and modelling, the use of real industrial data requires a highly flexible and versatile means of performing the necessary data pre-processing tasks along with the data analysis task. Python [184] is one such high-level, general-purpose, open-source programming language that fits these criteria. The large range of standard libraries provides numerous modules and packages for performing various tasks. Its popularity has fostered a large, supportive online community that provides extensive documentation, tutorials and examples which further increases its applicability for this case study. Furthermore, the simple and intuitive syntax makes it easy to learn for those with little or no programming experience. The library used to perform data manipulation and analysis is known as pandas. It provides data structures and functions needed for working with structured and tabular data such as the log files stored on the BMS. Other open-source programming languages have also been used such as R, it is not as versatile as Python given its focus on statistical computing and graphics [185].

As shown in Figure 5-6, the raw data in the BMS log files are not stored in a form ready for analysis. Therefore, several steps must be taken to transform the raw data in the log files into a dataset. A systematic approach is taken so that this process could be easily repeated for other AHUs and HVAC equipment data located in these files. Firstly, the semantic information in the data mapping file was read into the programming environment to locate the files that are relevant to the AHU under investigation. Secondly, the bespoke data collection process was reverse-engineered to create a time and value pair for each data point in the log file. The data for each feature was then merged with the other data recorded at the same time. A dataset was then created from the log files for the AHU. To further extend the dataset, the data available from the third-party company was also integrated. This data was exported in a CSV file format from the company's tool. This data was then presented in a tabular format which only required the bespoke names to be updated to align with the naming convention of the pre-processed log files. Lastly, an "update" command was then used which filled any missing data in the log file dataset with the data incoming from the third-party company dataset. This is a systematic process that may be repeated for other HVAC equipment simply by changing a single argument in the code regarding the system under study, given the log file has been appropriately recorded in the mapping file.

A snippet of some of the data preparation code is provided in Figure 5-8 and Figure 5-9 to reflect the steps taken in reverse engineering the log file to prepare the dataset. In brief, a loop is created to cycle through each log file, utilising a pandas dataframe (a tabular data structure), to transform the data into a wide dataframe (a single dataframe with each column representing one data point). An example of some of the steps includes; changing the timestamp format (line 69), changing the data type to DateTime (line 72) and reshaping the data from horizontal to vertical (line 76). This code in Figure 5-8 and Figure 5-9 reshapes the log files into a single tabular data structure. Further code has been developed to rename the data points based on the mapping file in Figure 5-7, prepare the data from the third-party data source, and then merge these two datasets. This is a time-consuming task without addressing the inevitable data quality issues within the dataset. Therefore, a structured approach to assessing this dataset is needed to ensure that issues are systematically addressed.

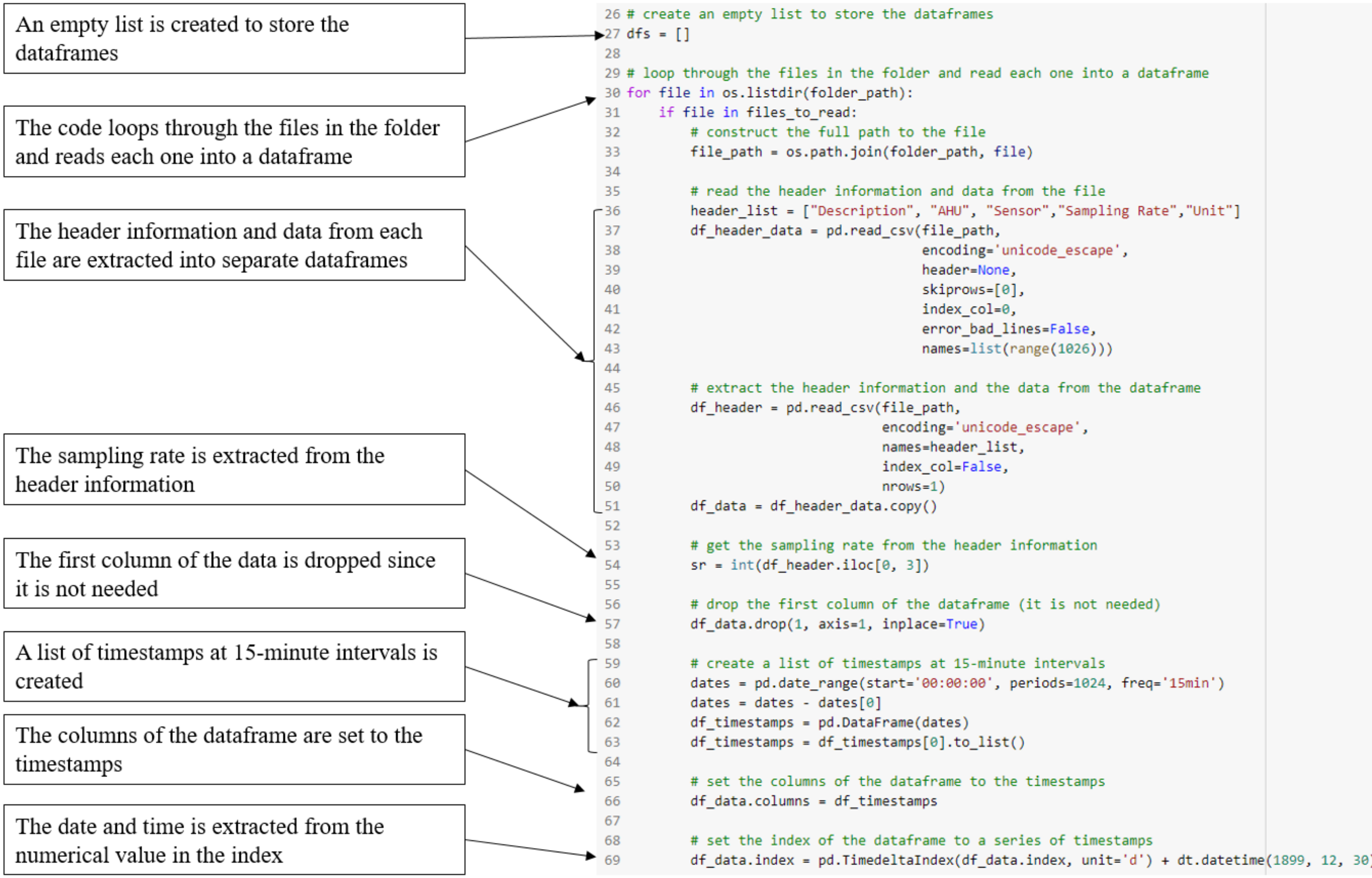


Figure 5-8 Data Preparation Code Snippet (Part 1)

The index of the dataframe is set to a series of timestamps

The index and columns of the dataframe are converted to datetime

The dataframe is reshaped so that the columns become the second level of the index

The index of the dataframe is converted to a single level by combining the values from the two levels of the index

The dataframe is reset to a default index

The columns of the dataframe are renamed

The values in the 'date' column are rounded to the nearest 15 minutes

The sensor and AHU information is added to the dataframe

Duplicates are removed and the list of all the dataframes is concatenated into one

```python

        # convert the index and columns of the dataframe to datetime
        df_data.index = pd.to_datetime(df_data.index)
        df_data.columns = pd.to_timedelta(df_data.columns)

        # reshape the dataframe so that the columns become the second level of the index
        df_data = df_data.stack()

        # convert the index of the dataframe to a single level by combining the values
        # from the two levels of the index
        df_data.index = df_data.index.get_level_values(0) + df_data.index.get_level_values(1)

        # reset the index of the dataframe
        df_data = df_data.reset_index()

        # rename the columns of the dataframe
        df_data.columns = ['date', 'val']

        # round the values in the 'date' column to the nearest 15 minutes
        df_data['date'] = df_data['date'].dt.round('15min')

        # rename the columns of the dataframe
        df_data = df_data.rename(columns={"date": "Datetime",
                                          "val": "Value"})

        # add the sensor and AHU information to the dataframe
        df_data["Sensor"] = df_header.iloc[0, 2]
        df_data["AHU"] = df_header.iloc[0, 1]
        df_data["Unit"] = df_header.iloc[0, 4]

        # drop duplicate rows from the dataframe
        df_data.drop_duplicates(inplace=True)

        # add the dataframe to the list of dataframes
        dfs.append(df_data)

# concatenate all of the dataframes in the list into a single dataframe
big_frame = pd.concat(dfs, axis=0, ignore_index=True)
```

Figure 5-9 Data Preparation Code Snippet (Part 2)

5.2.3 Data Assessment

Figure 5-10 Data Assessment phase of the IDAIC framework

In the third phase, the data is assessed to determine issues that need to be addressed before assessing the AHUs operation. While the data has been broadly determined to be sufficient in terms of time span and number of features, this phase delves into a more detailed assessment of data comprehensiveness (Broken), quality (Bad-quality) and consistency (Background). This is achieved through the steps that are summarised in Figure 5-11. Firstly, completeness is assessed by measuring the daily percentage availability of each feature. If the sensors required by the shallow knowledge expert rules to detect the visible faults in Table 5-3 are not recording data reliably, then a digitalisation plan is needed to acquire the necessary data. This would involve cataloguing the sensors that are missing and those that are not logging reliably so that a plan to capture this data may be made. Availability of over 85% is heuristically selected as the threshold to be considered sufficiently complete.

The further to the right of the decision tree in Figure 5-11, the greater the need for domain expertise. The quality of the data and the meaning of the data are determined through comparison with similar sensors and previous readings. A similar sensor is the term used to describe a data point that should closely align with another data point under a specified set of conditions. For example, if the valve control signal is at 0% then the air temperature before and after the coil should be similar in that no heating or cooling has altered its temperature. This logic is repeated for the air temperature sensors throughout the AHU. The results of the similar sensor comparison are then aligned to fault types. For example, if a continuous difference of 3°C is observed then a bias fault has occurred. The occurrence of a fault leads to the recommendation that sensors are manually checked and replaced or recalibrated. While sensor faults and transient state conditions have traditionally been a barrier to data-driven analysis, their identification in this phase enables the application of suitable data preparation techniques for the effective detection of energy-wasting faults. This phase is key to understanding the data in a structured, systematic and repeatable manner. The thresholds at each decision point in Figure 5-11 are heuristically selected using domain knowledge which is informed by the thresholds selected for APAR application [141] and the default values for sensor error in ASHRAE guideline 36 [66]. The thresholds used in this case study are presented in Table 5-6.

Table 5-6 Data Assessment Thresholds

Threshold	Value	Explanation
a	85%	Check that the percentage of non-null values exceeds 85%
b	95%	Check that the percentage of values that lie within the expected range exceeds 95%
c	0	Check that the difference from the previous value is not consecutively 0
d	1°C	Check that the absolute difference between similar sensors is less than 1°C
e	0.5°C	Check that the absolute rolling mean of the difference between similar sensors is less than 0.5°C
f	1°C	Check that the rolling standard deviation of the difference between similar sensors exceeds 1°C
g	±0.5°C, 0°C	Check that the difference between the previous rolling mean values is the same (within ±0.5°C), or is consecutively greater than or less than 0°C (increasing/decreasing).
h	0.5	Check if the sum of the steady-state detector exceeds 0.5
Window	4	For checks that include a rolling calculation or consecutive values, the window length is set to 4 values (1 hour)

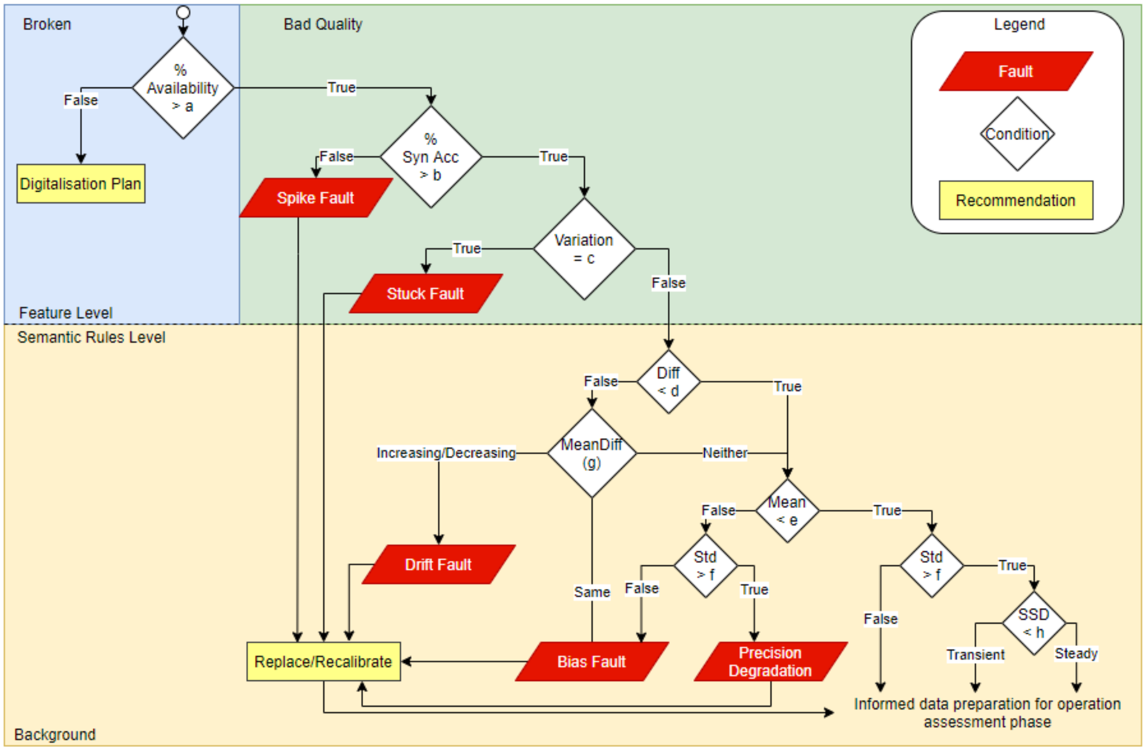


Figure 5-11 Data Assessment Decision Tree

The daily percentage availability and daily percentage syntactic accuracy, for each feature are visualised in the form of a heat map in Figure 5-12. The term syntactic accuracy refers to the form and structure of a data point. For example, the value returned for an outside air temperature sensor should be numerical and within a pre-determined range. A data point is considered syntactically accurate in this case study if the value lies within the expected range of values outlined in Table 5-1, as heuristically determined using domain expertise, for that sensor. The purpose of the syntactic accuracy assessment is to identify those sensor values that should be removed from the analysis as their readings are believed to be incorrect. While a general trend of improved availability and accuracy is seen over time, the energy meters on the supply and return fan (DaFanPower and RaFanPower) experience some communication issues after their installation in October 2020. Figure

5-12 (b) also shows many occurrences of the supply air and average zone temperature for zone 1 (ZoneDaTemp_1 and ZoneTemp_1) not aligning with the expected values.

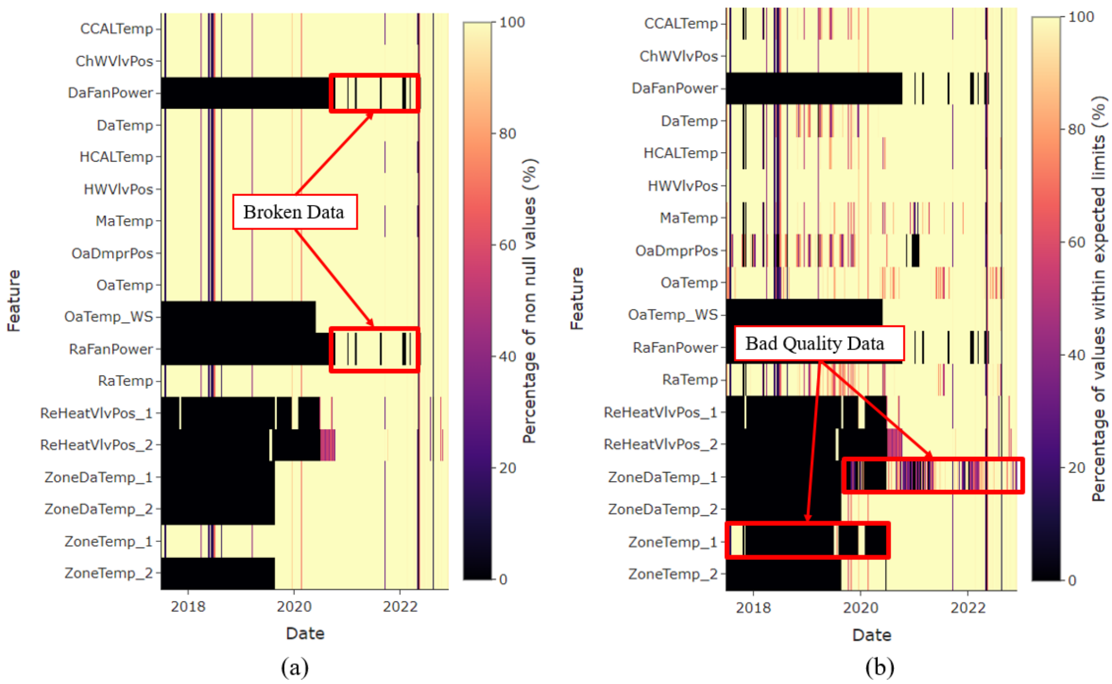


Figure 5-12 Daily percentage availability (a) and syntactically accurate values (b) for each feature

While syntactic accuracy identifies obvious occurrences of inaccurate readings, the semantic consistency assessment monitors the quality of the data more closely. The term semantic consistency refers to the coherence and meaningfulness of the data and is concerned with ensuring the data is logical and consistent. Therefore, statistical properties of the difference between similar sensors are assessed. For example, the outside air temperature (OaTemp) readings may be compared with the readings from a nearby weather station (OaTemp_WS). While the sensor readings in this example are likely to differ based on the distance between their respective locations, it offers a practical means of assessing the quality of the data.

The semantic rules level in Figure 5-11 firstly checks for a significant difference between similar sensors. As this instantaneous difference could be caused by a sensor fault or a component fault, the development of the inconsistency is analysed to understand the degradation pattern. Therefore, statistical properties of the difference such as the rolling mean (Mean), rolling standard deviation (Std) and the difference from the previous rolling mean value (MeanDiff) are used to classify sensor fault types. These properties are selected to provide a shallow knowledge form of the mathematical models used to express the different sensor fault types in Ullah Jan et al. [151]. Furthermore, the variability of the data is assessed to aid the data preparation phase. Therefore, the semantic rules level provides two key functions. Firstly, the statistical analysis of the difference between similar sensors over time provides greater insight than the application of an expert rule at one point in time. The in-depth analysis of the fault phenomenon aids industrial practitioners to distinguish sensor faults from component faults, which has been highlighted as an important future research task in the literature [18]. Secondly, as noted by House et al. [141], expert rules are not restricted to steady-state conditions but the likelihood of false alarms increases when the system is in a transient state. The integration of a steady-state detector (SSD), utilised by Lee et al. [155], improves the applicability of expert rules by identifying transient state periods that require the use of smoothening techniques such as exponentially weighted moving averages.

An example of the outside air temperature assessment is shown in Figure 5-13. At 3:00 am on the 29th of January 2021, OaTemp begins to deviate from OaTemp_WS. At 4:00 am, a drift fault is detected because there is a continuously increasing discrepancy between these sensors. The rate of increase of the discrepancy between these sensors then approached zero, which triggered a bias fault at 4:30 am. The difference between the two sensors did not remain constant, therefore another drift fault is detected at 6:45 am. This cycle continued until 2:45 am the following day, at which point the sensor readings converged and the status returned to normal. Further investigation of this discrepancy uncovered OaTemp readings greater than 40°C, which is known to be inaccurate considering the geographical location of this site.

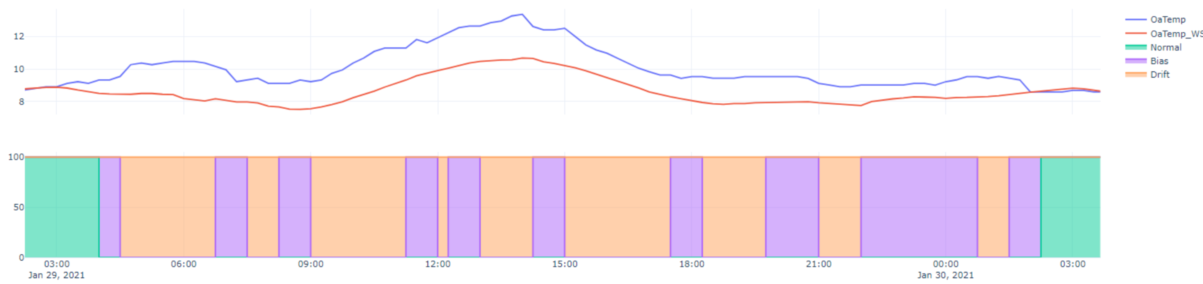


Figure 5-13 Example of drift and bias sensor fault detection for the outside air temperature sensor

This process is repeated for other sensors as in Table 5-7. As it may be difficult to determine which of the two sensors caused the discrepancy, each temperature sensor is compared to at least one other sensor, where possible. Conditions are included where necessary to ensure that the sensors used in the comparison are measuring the same physical phenomena. For example, in SC_3, the hot water valve control signal must be fully closed to ensure there has been no heating of the air between these sensors. To reduce the effect of residual heat in the coil, each condition must be satisfied for a reasonable period, heuristically selected as one hour or four consecutive values. While this is a larger threshold in comparison to the 12-minute time frame selected for steady-state conditions by Raftery et al [186], the 15-min granularity of this dataset means that the selection of only a few data points leads to a long time period. The assumption is that four consecutive fully closed values would indicate a period without heating, reducing the chance of an unseen valve opening between timestamps which may occur if a smaller time frame was selected.

Table 5-7 Semantic Rules

Rule Number	Comparison	Condition
SC_1	OaTemp vs OaTemp_WS	None
SC_2	MaTemp vs OaTemp	OaDmprPos = 100%
SC_3	HCALTemp vs MaTemp	HWVvPos = 0%
SC_4	CCALTemp vs HCALTemp	ChWVlvPos = 0%
SC_5	DaTemp vs CCALTemp	None
SC_6	CCALTemp vs MaTemp	HWVvPos = 0% AND ChWVlvPos = 0%
SC_7	DaTemp vs MaTemp	HWVvPos = 0% AND ChWVlvPos = 0%
SC_8	ZoneDaTemp_1 vs DaTemp	ReHeatVlvPos_1 = 0%
SC_9	ZoneDaTemp_2 vs DaTemp	ReHeatVlvPos_2 = 0%

Further investigation of the semantic rules, regarding the rolling mean and standard deviation, give a better indication of the accuracy of these sensors. Precision degradation may be monitored by analysing the changes in the mean and standard deviation. High variation in these statistical measures (standard deviation exceeding 1°C) uncovered that the AHU was not operating in a steady state. To quantitatively get more insight into the background of the data, a steady-state detector (SSD) proposed by Lee et al. [155] is applied to the control signals (outside air damper control signal, hot water valve control signal and chilled water valve signal). When the data is available, the sum of the three SSD values must be below the heuristically selected threshold (h in Table 5-6) for the AHU to be considered in steady-state. Otherwise, it is classified as a transient state. The results graphed in Figure 5-14 show a general decrease in the time spent in steady-state conditions at the end of 2019. Through discussions with the onsite facilities team at the stakeholder meetings, this time was determined to have aligned with a change in the operation of the AHU. Due to the COVID-19 pandemic, the operation of the AHU was changed to allow the maximum quantity of fresh air into the building as a measure to reduce the likelihood of COVID-19 infection. In short, this essentially meant that a recirculating AHU was operating as a full fresh-air AHU. Consequently, established knowledge-based approaches that were to be applied in the operation assessment phase would likely require an extension to take account of this novel operation. This is a key finding of the data assessment phase, along with the identification of sensor faults which are summarised in Table 5-8 in the next subsection.

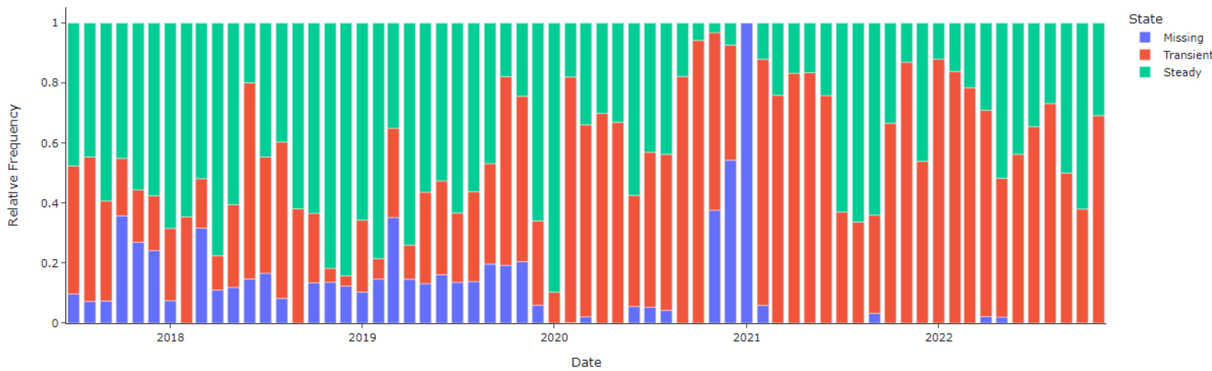


Figure 5-14 Relative frequency of data background

5.2.4 Data Preparation

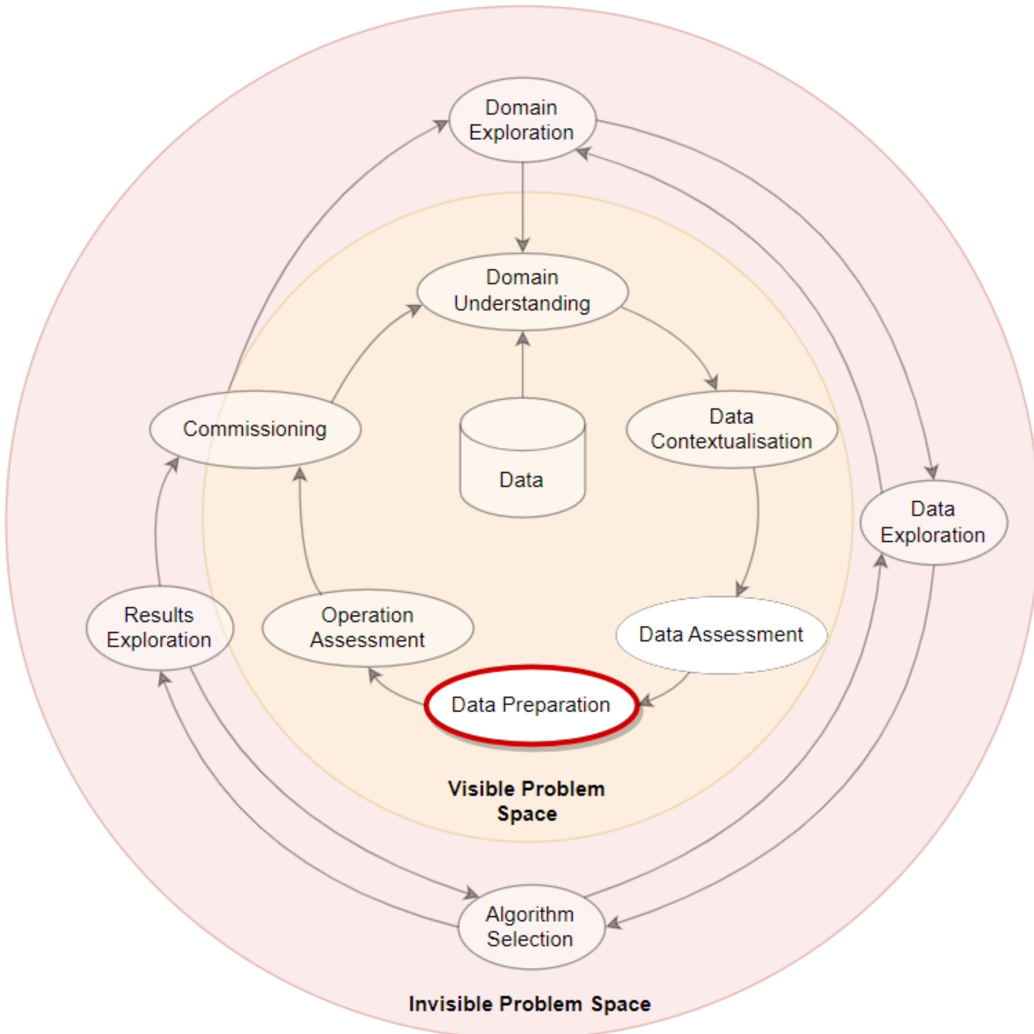


Figure 5-15 Data Preparation phase of the IDAIC framework

In the fourth phase, the data is prepared for the knowledge-based FDD method. The incomplete, inaccurate and inconsistent data issues identified in the data assessment phase were then addressed. Given that the data cleaning process could negatively impact the assessment of the AHUs operation in later phases, the mitigation of the observed issues was discussed and decided upon at the stakeholder meetings. This ensures that the issues identified in Table 5-8 are systematically addressed by incorporating domain expertise. For example, the occurrence of many instances of drift, bias and precision degradation faults for the outside air temperature sensor, led to the decision to replace this feature with the values from the nearby weather station. This is not an ideal solution, but a practical one that enabled the application of retrospective knowledge-based FDD in the presence of a faulty sensor. Recalibration or replacement of the outside air temperature sensor is a task to be completed in the commissioning phase.

Regarding the background issue, two main observations must be addressed. Firstly, the outside air damper control signal manually set to 100% open meant that the AHU was operating in a "new" control mode not identified in the FDS. This "COVID_Operation" requires updates to existing rulesets to apply FDD to AHUs operating under these constraints. Secondly, the high proportion of transient operation increases the difficulty of applying FDD techniques. Therefore, additional FDD rules need to be developed for COVID_Operation and transient conditions need to be smoothened using statistical techniques. The operation assessment phase will therefore require the development of a new FDD method to account for these issues.

Table 5-8 Example of Data Preparation Tasks

	Challenge	Observation	Mitigation
1	Heterogeneous Data Sources	DateTime not at 15-min intervals from July 2020 to Oct 20	Removed data not at a multiple of 15mins
2	Erroneous timestamp	Data for 2015 mistakenly returned	Data before June 2017 removed from the analysis
3	Erroneous data	DaFanPower granularity greater than two decimal places	The full dataset rounded to two decimal places
4	Missing Data	RaFanPower and DaFanPower missing Oct 2020	Feature removed from the analysis before this date
5	Missing Data	Most features missing from 30/04/22 to 16/05/22	This occurred due to a planned shutdown. This period is not analysed.
6	Spike	ZoneDaTemp_1 higher than expected	Further exploration is required in invisible problem space regarding reheat to this zone
7	Spike	ZoneTemp_1 returns mostly +500 or -500 from December 2017 to June 2020	Assumed to be at the set point during this period to enable the FDD application
8	Spike	Most features contain outliers	Left unchanged for closer FDD inspection

	Challenge	Observation	Mitigation
9	Spike	DaTemp reading 1000°C on 02/03/21	These instances are removed from the dataset
10	Spike	OaTemp reading above 600°C on 08/10/22	Removed and replaced with OaTemp_WS
11	Stuck	Control signals read 0 in Nov 2020 and Dec 2020 to Feb 2021	OaDmprPos manually override to 100% in this period. Valve control signals were removed.
12	Drift	Faulty OaTemp	Replaced with OaTemp_WS
13	Stuck	OaDmprPos stuck at 100% open	Manual override. Upgrade to knowledge-based FDD technique required to account for a novel mode of operation
14	Stuck	RaFanPower & DaFanPower often become stuck at a value	Communications issue. Assumed to be true given the low variability of the data on inspection. However, a step change occurred when one of the issues was resolved which indicates the assumption may be poor during this period (May-July 2022)
15	Bias	DaTemp > CCALTemp	An expected, but inconsistent, rise across the supply fan. Mean and standard deviation used to inform the development of the FDD method
16	Bias	DaTemp > ZoneDaTemp_2	Similar to the bias between DaTemp and CCALTemp. DaTemp must be recalibrated.
17	Precision Degradation	Air temperature sensor readings are often erratic due to unstable control	Transients are smoothened using an exponentially weighted moving average as in House et al. [32]

5.2.5 Operation Assessment

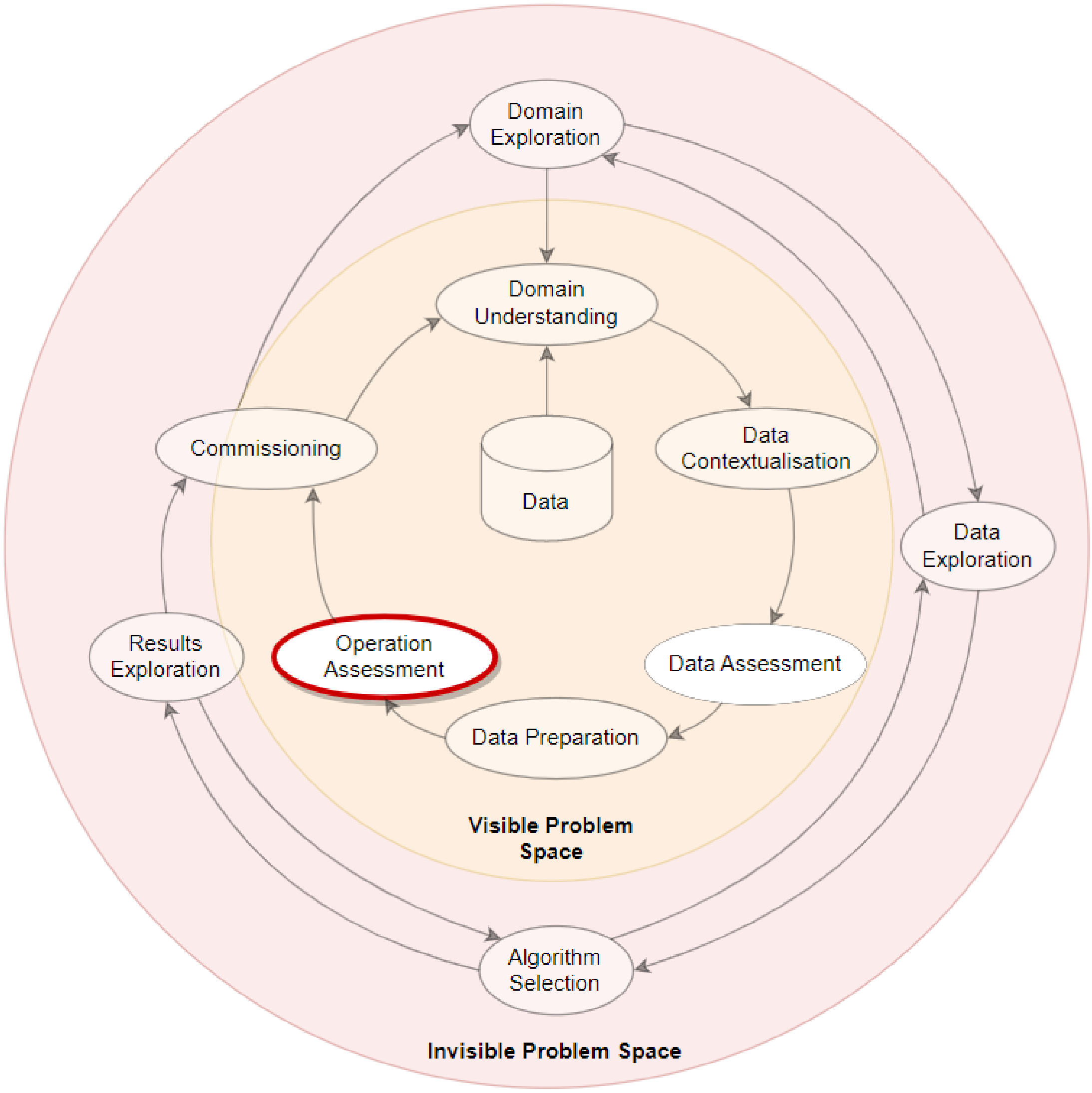


Figure 5-16 Operation Assessment phase of the IDAIC framework

The goal of the Operation Assessment is to identify the visible faults in Table 5-3. Given the presence of data issues, the applicability of FDD methods may reduce which in turn reduces the number of faults which may be correctly detected. Given the pervasiveness of the data issues, a data-driven method for fault diagnosis is not currently suitable and a highly interpretable technique is required. The AHU performance assessment rule set (APAR) [141] was selected as the most appropriate method given its status as a research-grade FDD algorithm [138] and its computational simplicity which increases its practical applicability [183]. APAR was extended to account for COVID_Operation by defining new modes of operation; mechanical heating with 100% outdoor air (1_c), free cooling with 100% outdoor air (2_c), mechanical cooling with 100% outdoor air (3_c), and simultaneous mechanical heating and cooling with 100% outdoor air (4_c). The rules associated with modes 1, 2 and 3 from the APAR ruleset are replicated in modes 1_c, 2_c and 3_c respectively. Mode 4_c is considered a fault in this case study. While mode 3_c may be considered the same as mode 3, the mode is repeated for the sake of completeness. The full ruleset is provided in Figure 5-17, with a summary of the novel rules as follows;

In mechanical heating with 100% outdoor air constraint mode (Mode 1_c), Rule 1_c indicates that the supply air temperature is less than the mixed air temperature, accounting for an air temperature increase across the supply fan and air temperature sensor error. Assuming that the air temperature sensors are accurate based on the data assessment phase, the triggering of this rule would indicate a possible leaking cooling coil valve or an issue with the hot water system such as pump failure. Given that Rule 2 is related to the fraction of outdoor air entering the AHU, it should be high given the ventilation purposes of the COVID_Operation (Mode 2_c). Rule 2 was originally intended to detect if the fraction is too high or too low, therefore it should be updated to only trigger if it is too low in this mode. The direction of the inequality is reversed so that the true condition, or fault condition, indicates a low outdoor air fraction. Additionally, the threshold parameter accounting for errors related to airflows, εf (0.3), is updated to 1-εf (0.7), so that only a high outdoor air fraction is considered not a fault. Rule 3_c and Rule 4_c are related to rules 3 and 4 which indicate that the heating coil valve is saturated at the fully open position. This is most likely to occur in the winter season given that the heating coil is the only means of heating the outdoor air that is passing through the AHU.

It is unlikely that a prolonged period is spent in the cooling with 100% outdoor air constraint mode (Mode 2_c). The reason for this is that the AHU has no means of controlling the supply air temperature in this mode. This mode should only occur if the outdoor air temperature is the same as the required supply air temperature, or within the appropriate deadband limits, as there is no mechanical heating or cooling. Similar to the

APAR rules, Rules 5_c, 6_c and 7_c indicate an inconsistency between the air temperature sensors.

As previously mentioned, the mechanical cooling with 100% outdoor air constraint mode (Mode 3_c) is the same as APAR mode 3. As is the nature of the original APAR rules, there is repetition between past rules as follows; Rule 8_c and Rule 5_c, Rule 11_c and Rule 1_c, Rule 12_c and Rule 6_c, Rule 13_c and Rule 3_c and finally, Rule 14_c and Rule 4_c. There are two rules unaccounted for, Rule 9_c and Rule 10_c. Rule 9 is not extended to Rule 9_c because the indication of the outdoor air temperature being too high for this mode does not apply. Ordinarily, Rule 9 would indicate that mechanical cooling with minimum outdoor air (Mode 4) should be in operation but this is not possible given the 100% outdoor air constraint for COVID_Operation. Therefore, Rule 9_c is not defined. Meanwhile, Rule 10_c indicates the possibility of a more critical fault in COVID_Operation than its Rule 10 origin. Similar to Rule 2_c, Rule 10_c indicates a difference between the mixed air temperature sensor and the outdoor air temperature sensor. While this may still be caused by an air temperature sensor error, it may also indicate the leakage of return air into the mixing chamber. This is considered a fault in the COVID_Operation given the ventilation implication.

Finally, the simultaneous mechanical heating and cooling with 100% outdoor air constraint (Mode 4_c) is considered a fault mode for this AHU. From an energy consumption perspective, it is a key fault to quickly identify and rectify given the waste associated with heating and cooling the outdoor air at the same time. The rule, Rule 22_c, originates from Rule 22 in the unknown occupied modes (Mode 5) of APAR. It is the only rule from Mode 5 that is extended to COVID_Operation given the use of damper positions in Rules 21, 23 and 24. Furthermore, the rules pertaining to all occupied modes, specifically rules 25-28, remain applicable.

Mode	Rule #	Rule Expression (true implies existence of a fault)
Heating (Mode 1)	1	$Tsa < Tma + \Delta Tsf - \varepsilon t$
	2	For $\lvert Tra - Toa \rvert \geq \Delta Tmin$: $\lvert Qoa/Qsa - (Qoa/Qsa)min \rvert > \varepsilon f$
	3	$\lvert uhc - 1 \rvert \leq \varepsilon hc$ and $Tsa,s - Tsa \geq \varepsilon t$
	4	$\lvert uhc - 1 \rvert \leq \varepsilon hc$
Cooling with Outdoor Air (Mode 2)	5	$Toa > Tsa,s - \Delta Tsf + \varepsilon t$
	6	$Tsa > Tra - \Delta Trf + \varepsilon t$
	7	$\lvert Tsa - \Delta Tsf - Tma \rvert > \varepsilon t$
Mechanical Cooling with 100% Outdoor Air (Mode 3)	8	$Toa < Tsa,s - \Delta Tsf - \varepsilon t$
	9	$Toa > Tco + \varepsilon t$
	10	$\lvert Toa - Tma \rvert > \varepsilon t$
	11	$Tsa > Tma + \Delta Tsf + \varepsilon t$
	12	$Tsa > Tra - \Delta Trf + \varepsilon t$
	13	$\lvert ucc - 1 \rvert \leq \varepsilon cc$ and $Tsa - Tsa,s \geq \varepsilon t$
	14	$\lvert ucc - 1 \rvert \leq \varepsilon cc$
Mechanical Cooling with Minimum Outdoor Air (Mode 4)	15	$Toa < Tco - \varepsilon t$
	16	$Tsa > Tma + \Delta Tsf + \varepsilon$
	17	$Tsa > Tra - \Delta Trf + \varepsilon t$
	18	For $\lvert Tra - Toa \rvert \geq \Delta Tmin$: $\lvert Qoa/Qsa - (Qoa/Qsa)min \rvert > \varepsilon f$
	19	$\lvert ucc - 1 \rvert \leq \varepsilon cc$ and $Tsa - Tsa,s \geq \varepsilon t$
	20	$\lvert ucc - 1 \rvert \leq \varepsilon cc$
Unknown Occupied Modes (Mode 5)	21	$ucc > \varepsilon cc$ and $uhc > \varepsilon hc$ and $\varepsilon d < ud < 1 - \varepsilon d$
	22	$uhc > \varepsilon hc$ and $ucc > \varepsilon cc$
	23	$uhc > \varepsilon hc$ and $ud > \varepsilon d$
	24	$\varepsilon d < ud < 1 - \varepsilon d$ and $ucc > \varepsilon cc$
All Occupied Modes (Mode 1, 2, 3, 4, or 5)	25	$\lvert Tsa - Tsa,s \rvert > \varepsilon t$
	26	$Tma < min(Tra , Toa) - \varepsilon t$
	27	$Tma > max(Tra , Toa) + \varepsilon t$
	28	Number of mode transitions per hour > MTmax
Heating with 100% Outdoor Air Constraint (Mode 1_c)	1_c	$Tsa < Tma + \Delta Tsf - \varepsilon t$
	2_c	For $\lvert Tra - Toa \rvert \geq \Delta Tmin$: $\lvert Qoa/Qsa - (Qoa/Qsa)min \rvert < 1-\varepsilon f$
	3_c	$\lvert uhc - 1 \rvert \leq \varepsilon hc$ and $Tsa,s - Tsa \geq \varepsilon t$
	4_c	$\lvert uhc - 1 \rvert \leq \varepsilon hc$
Cooling with 100% Outdoor Air Constraint (Mode 2_c)	5_c	$Toa > Tsa,s - \Delta Tsf + \varepsilon t$
	6_c	$Tsa > Tra - \Delta Trf + \varepsilon t$
	7_c	$\lvert Tsa - \Delta Tsf - Tma \rvert > \varepsilon t$
Mechanical Cooling with 100% Outdoor Air Constraint (Mode 3_c)	8_c	$Toa < Tsa,s - \Delta Tsf - \varepsilon t$
	10_c	$\lvert Toa - Tma \rvert > \varepsilon t$
	11_c	$Tsa > Tma + \Delta Tsf + \varepsilon t$
	12_c	$Tsa > Tra - \Delta Trf + \varepsilon t$
	13_c	$\lvert ucc - 1 \rvert \leq \varepsilon cc$ and $Tsa - Tsa,s \geq \varepsilon t$
	14_c	$\lvert ucc - 1 \rvert \leq \varepsilon cc$
Simultaneous Mechanical Heating and Cooling with 100% Outdoor Air Constraint (Mode 4_c)	22_c	uhc >!=0 and ucc != 0

Figure 5-17 Extended APAR ruleset with extended rules highlighted in red

Table 5-9 Definition of Nomenclature in extended APAR ruleset

Symbol	Meaning
MTmax	maximum number of mode changes per hour
Tsa	supply air temperature
Tma	mixed air temperature
Tra	return air temperature
Toa	outdoor air temperature
Tco	changeover air temperature for switching between Modes 3 and 4
Tsa,s	supply air temperature set point
ΔTsf	temperature rise across the supply fan
ΔTrf	temperature rise across the return fan
ΔTmin	threshold on the minimum temperature difference between the return and outdoor air
Qoa/Qsa	outdoor air fraction = (Tma - Tra)/(Toa - Tra)
(Qoa/Qsa)min	threshold on the minimum outdoor air fraction
uhc	normalized heating coil valve control signal [0,1] where uhc = 0 indicates the valve is closed and uhc = 1 indicates it is 100 % open
ucc	normalized cooling coil valve control signal [0,1] where ucc = 0 indicates the valve is closed and ucc = 1 indicates it is 100 % open
ud	normalized mixing box damper control signal [0,1] where ud = 0 indicates the outdoor air damper is closed and ud = 1 indicates it is 100 % open
εt	threshold for errors in temperature measurements
εf	threshold parameter accounting for errors related to airflows (function of uncertainties in temperature measurements)
εhc	threshold parameter for the heating coil valve control signal
εcc	threshold parameter for the cooling coil valve control signal
εd	threshold parameter for the mixing box damper control signal

Using the possible explanations of faults from House et al. [141], the rules that may trigger as a result of the occurrence of the faults are listed in Table 5-3. As the data assessment phase is focused on detecting the occurrence of potential sensor faults, the diagnosis of triggered APAR rules is aided. For example, the violation of rule 1 could be caused by one of 2 sensor faults or one of 6 component faults. If the data quality is determined to be sufficient, then it is likely to be one of the remaining 6 component faults. To determine the specific component, analysis of both the data trends and other triggered faults is useful. This enables the likely root cause of the fault to be prioritised to increase the effectiveness of maintenance actions. Given that only the features required for APAR have been used to analyse the operation, incorporating more features may be beneficial for the diagnosis task. For example, the availability of data for the hot and chilled water, such as temperature and flow rate, may potentially differentiate between water pump failure and valve failure. These features are not available in this case study, therefore, the recommendation may be to manually open the valve and check the temperature difference across the coil. However, further analysis could aid better detection and diagnosis.

While various methods exist to deal with fault diagnosis, such as fault isolation algorithms based on decision trees [187], certain faults may pose challenges in identification due to the self-correcting nature of closed-loop control or the influence of feedback control [188]. The IDAIC framework is designed to allow flexibility so that an appropriate method may be selected to solve a particular problem. Given the pervasiveness of the data issues in this case study, the issue of hidden faults is addressed by isolating sensor, component and control faults from each other. Sensor faults are analysed by comparing the values of similar sensors. For example, SC_2 and SC_3 in Table 5-7 may be analysed to isolate any issues with the mixed air temperature sensor. The visible component faults are detected using APAR in this phase, while analysis of invisible component faults is left for the invisible problem-solving activity. Lastly, control faults are detected in both the Data Assessment phase using the steady-state detector for transient control, and also in this phase using APAR rules 21 to 24 to identify controller logic errors in unknown modes of operation. The difference between the approach in this framework and other existing methods is that data assessment, knowledge-based FDD and data-driven FDD are used together to give a holistic view of the operation of the AHU at the sensor, component and control level. This gives decision makers greater understanding which they may use to prioritise maintenance actions.

While the goal of the operation assessment is to detect visible faults, some invisible faults may be detected using the APAR ruleset. However, it may be assumed that invisible faults will only trigger APAR rules when the severity of the fault is high as the abnormal characteristic generated must exceed a pre-defined threshold. In an ideal proactive maintenance approach, the fault would be addressed before the occurrence of a high-severity fault, or at least monitored before then. Therefore, further data-driven analysis is required to ensure lower severity, yet energy-wasting, faults are identified early.

The percentage of time spent in a mechanical heating or cooling mode each year is compared to the time spent in free cooling (Mode 2) or an unknown mode of operation in

Figure 5-18. The most frequent mode of operation is free cooling, which is expected given that the manufacturing facility is located in a temperate oceanic climate whereby outdoor air generally provides sufficient cooling. A significant increase in the time spent in mechanical heating and cooling is observed in the years 2020 and 2021 due to the 100% outdoor air constraint. While the applicability of the APAR ruleset has been improved by extending the ruleset for COVID_Operation and performing informed pre-processing on the data, many faults are triggered as shown in Figure 5-19. On inspection, one of the most commonly reoccurring faults is Rule 28 which compares the number of mode transitions per hour with a heuristically selected threshold. Given that the data is at 15-min intervals, the threshold value of 6 suggested by House et al. [141] would never be reached because only four data points are available for one hour. Therefore the threshold was changed and the rule triggers only when there are 2 or more mode transitions in one hour. This controller tuning error has a significant impact on the number of false alarms in this dataset. For example, Rule 7 is the most frequent fault in the dataset which is triggered when there is an inconsistency between the supply air temperature and mixed air temperature in the free cooling mode (Mode 2). In the case of continuous cycling between the free cooling mode and mechanical cooling with 100% outdoor air mode (Mode 3), Rule 7 was triggered frequently because there was residual cooling occurring when the valve was shut. While smoothening techniques have been applied to the air temperature sensors in transient states, the application of these techniques to the control signals causes difficulty in mode identification. Therefore, the control logic should be reviewed to reduce the number of mode transitions so that unnecessary maintenance activity is avoided.

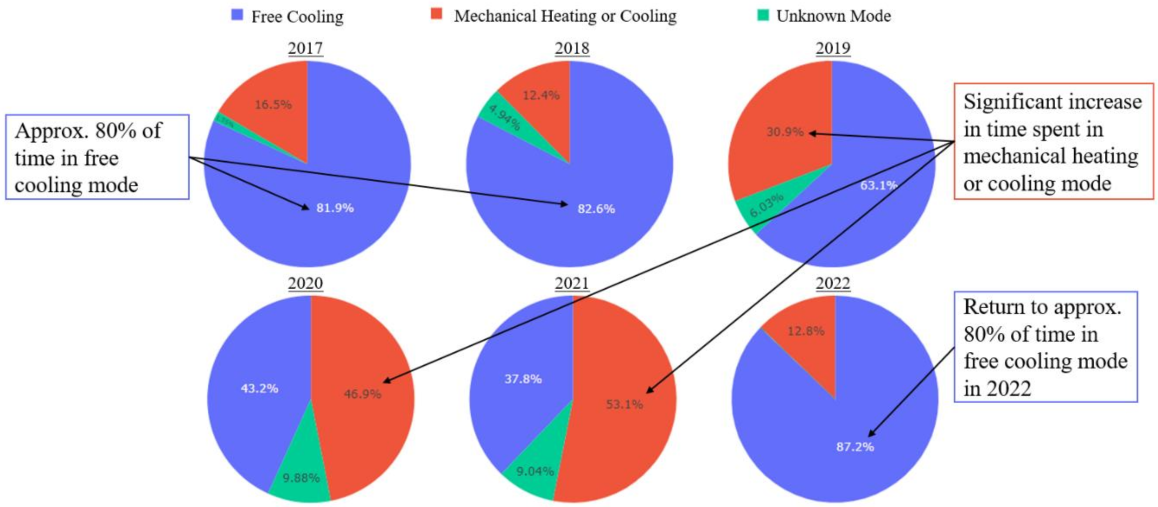


Figure 5-18 Time spent in a mechanical heating or cooling mode vs free cooling mode by year

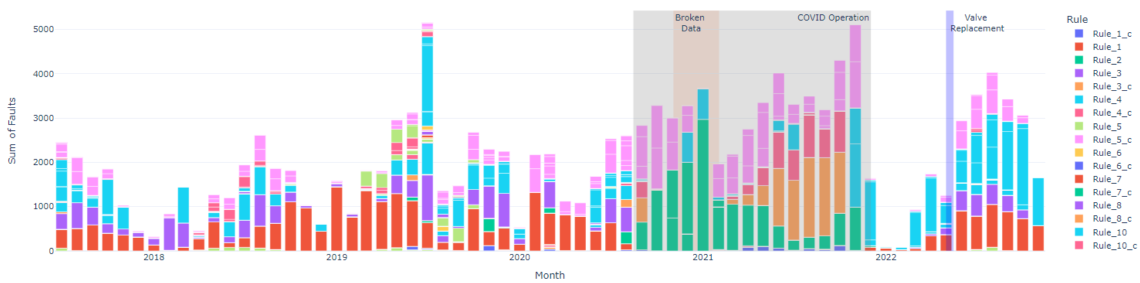


Figure 5-19 Total monthly sums of visible faults detected using the amended APAR ruleset

5.2.6 Commissioning

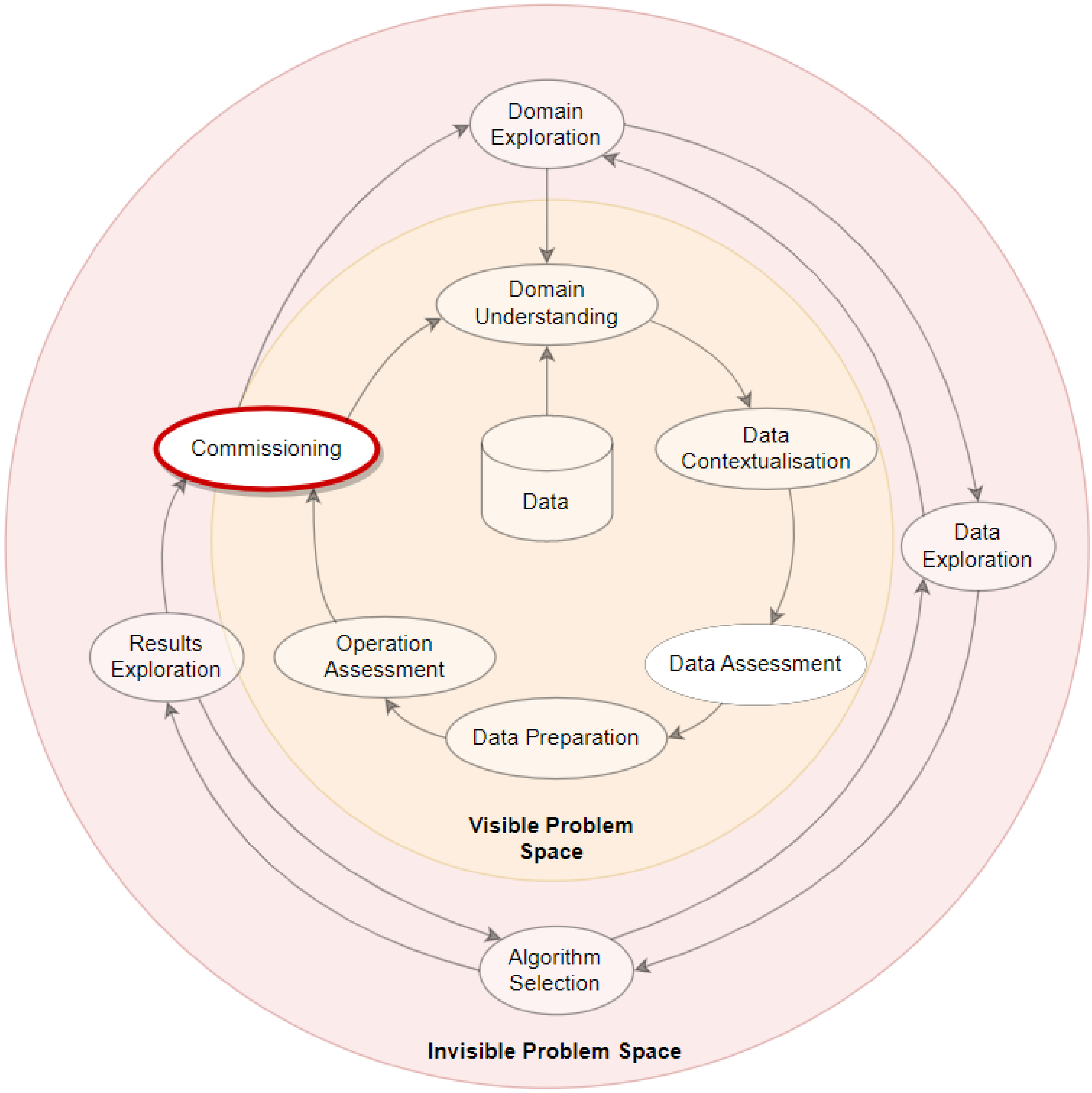


Figure 5-20 Commissioning phase of the IDAIC framework

The findings from the Data Assessment and Operation Assessment phases suggest that several maintenance actions are required. Firstly, the control strategy of the AHU needs to be reviewed. The AHU spends a significant proportion of its time in a transient state which causes difficulty in the analysis of the unit. Secondly, the outside air temperature sensor needs to be manually checked and either replaced or recalibrated. Thirdly, the AHU components and actuators should be manually checked and replaced where necessary to ensure the AHU is returned to its "as-designed" state. Therefore, the visible problem-solving activity of the IDAIC framework resulted in the identification of five key issues with only three main maintenance activities taking place as in Table 5-10. Firstly, on 02/12/2021, the control strategy was reviewed and the COVID-19 100% fresh-air operation was discontinued. Secondly, on the same date, large temperature differences across the coils led to the internal air temperature sensors (G, I, K, and M in Figure 5-4) being calibrated. Thirdly, with regards to physically inspecting the temperature differences across closed coils, for the same large temperature difference issue, the cooling and heating control valves were replaced during a planned shutdown on 01/05/2022. The actions left for future work, which would also not have been identified without the IDAIC framework given the issues in Table 5-8, include resolving the transient state issue and resolving the outside air temperature sensor calibration issue. This phase concludes the visible problem-solving activity.

Table 5-10 Overview of identified faults and associated maintenance action

Issue	**Recommendation**	**Status**
Discontinue COVID_Operation	Review control strategy	Completed on 02/12/2021
Large temperature differences across coils	Recalibrate temperature internal air temperature sensors	Completed on 02/12/2021
	Physically check temperature differences across closed coils	Completed and valves replaced on 01/05/2022
Faulty outside air temperature sensor	Replace/Recalibrate	Not completed
Transient control	Review control strategy	Not completed

An overview of the main outcomes in each phase of the visible problem-solving activity is provided in Figure 5-21. The total time to conduct these six phases was approximately 18 months. The analysis on this case study site was time-consuming given the siloed nature of the data and the lack of a detailed history of the AHU regarding past maintenance actions. This led to many data issues identified in the data assessment phase that contribute to the data gap, along with operational faults identified in the operation assessment phase which contribute to the operational gap. While some of these issues have been rectified in the commissioning phase, the faulty outside air temperature sensor and the transient control are unresolved, inhibiting an accurate fault-free baseline to be obtained. These unresolved visible problems act as an indicator that the value of data-driven methods may not be realised as they may mask the occurrence of subtle invisible problems. However, while accurate fault-free baseline operation cannot be obtained, the novel COVID_Operation may be leveraged to progress to the invisible problem-solving phases. As the COVID_Operation was not returned to normal operation for this AHU when the decision was made to do so, the continuation may be considered an invisible fault. While this is not a typical subtle degradation type fault, it is an example of an unnoticed fault which the invisible problems are. The COVID_Operation may then be used to label the dataset to prove the concept of invisible problem-solving given the significant deviation from normal operation this fault would cause.

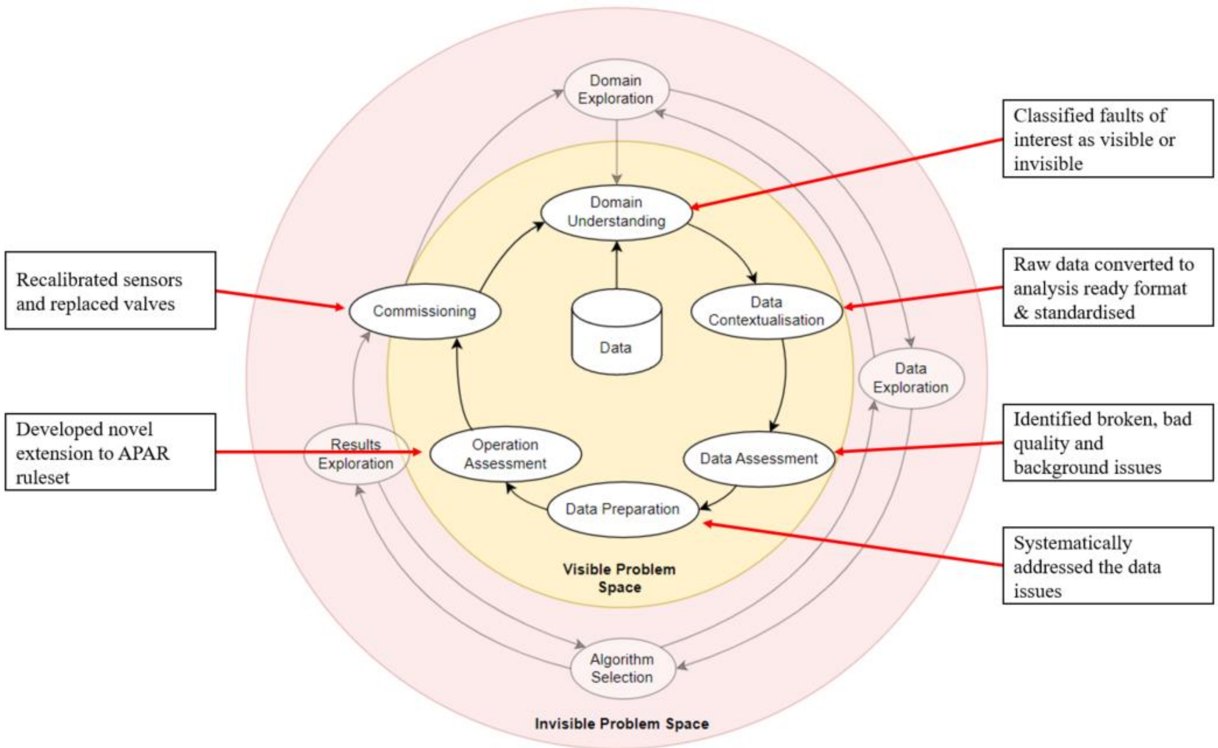


Figure 5-21 Main outcomes of each phase in the visible problem-solving space

5.3 Implementation of IDAIC for Invisible Problem-Solving

Invisible problems are those faults that are not easily identifiable. They differ from the goal-based analysis of visible faults as they require exploratory-based analysis. An invisible fault is typically characterised by subtle deviations from normality that indicate the occurrence of degradation. Therefore, the presence of visible faults may mask or hide the presence of an invisible fault. For example, an AHU operating in a transient state as a result of a control fault may cause significant differences between temperature values that are recorded every 15 minutes. The significant temperature difference caused by the control fault may then conceal the subtle temperature differences that may be caused as a result of a leaking heating or cooling coil valve. The case study site has not resolved all the issues identified in the visible problem-solving loop, therefore full implementation of the invisible problem-solving loop is not possible. However, some preliminary and practical data-driven analysis may be conducted in the exploratory phases.

One such example would be the analysis of heating coil fouling. The deposit of dust or other debris on the coil could lead to a reduction in the heating efficiency of the AHU. This is a typical invisible problem in that loss of efficiency occurs for a significant portion of time before it is identified by APAR through the system becoming out of control (Rule 4), or the valve becoming saturated at the fully open position and not reaching the air temperature set point (Rule 3). The phases necessary to detect this type of problem are discussed in the following subsection 5.3. Ideally, these phases follow from the rectification of the issues identified in subsection 5.2 so that the presence of known problems does not influence the exploration of abnormalities. In the heating coil fouling example, a clustering algorithm which uses key features relating to the heating coil, as defined using domain expertise, may be useful to identify changes in the heating efficiency of the coil. However, before an algorithm is applied, the problem must be understood. In the case of detecting heating coil fouling, the dynamics of the hot water system, such as flow rate and inlet temperature, would have a significant effect on the usefulness of the clustering algorithm. Although, its application may lead to the identification of abnormal behaviour which could be heating coil fouling, change in hot water supply temperature, hot water circulating pump fault or otherwise. Therefore, exploring the domain and the data is key to effectively using this information to prioritise maintenance actions.

5.3.1 Domain Exploration

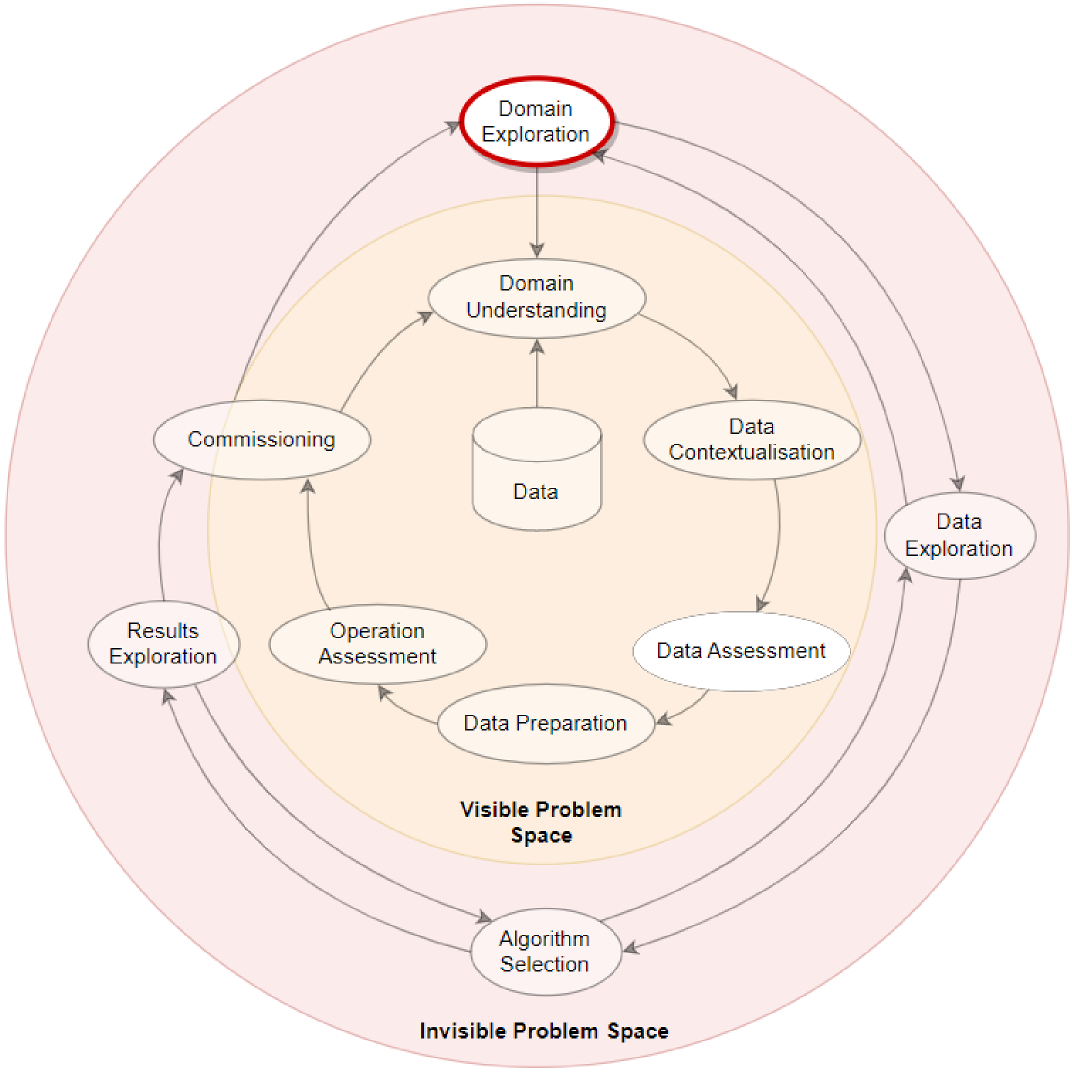


Figure 5-22 Domain Exploration phase of the IDAIC framework

The completion of the visible problem-solving loop, via the commissioning step, enabled the exploratory activity of the invisible problem space to begin. The goal of this phase is to determine areas of further exploration that may lead to uncovering invisible faults. In the Operation Assessment, the control strategy of this AHU made it difficult to apply expert rule sets given its transient nature. While smoothening techniques were applied which incorporate previous sensor readings, there is potential to uncover greater insight by analysing the clusters and patterns that develop over a longer period.

The faults of interest in this case study are those that lead to a reduction in energy efficiency. Therefore, the total energy consumed by the unit and abnormal deviations from the expected energy consumption are of interest. Additionally, monitoring the control of the AHU and analysing the performance degradation of the mechanical components are of interest. As discussed in the literature review in Chapter 2, data-driven methods in literature generally require a fault-free baseline which is then compared to incoming data to detect faults by identifying significant differences from the baseline. Given that a fault-free baseline cannot be obtained until all the visible problems are solved, the application of an algorithm would lead to a false sense of readiness for data-driven approaches, potentially leading to missed faults that would undermine the approach in future analysis. Therefore, the analysis should remain exploratory in nature, with domain knowledge used to interpret and guide the analysis. A highly interpretable approach that analyses the AHU on three levels, energy consumption, control and degradation, is therefore needed.

In similar work, Gao et al. [189] proposed a three-level data-driven fault detection solution as in Figure 5-23. The first level analysed the daily energy consumption, the second level analysed the precision and stability of the AHU and the third level analysed the residuals from actual and predicted values to identify faulty components. While the high-quality ASHRAE 1312-RP [137] dataset and data from an office building are used by Gao et al. [189], the limitations of the dataset in this case study require a more rudimentary approach. For example, daily cooling energy cannot be accurately quantified using the available data. This dataset is missing key features, such as water temperatures and flow rates as identified by Hosamo et al. [190], which limits the applicability of many FDD methods. However, the temperature decrease across the coils may be used as a proxy indicator. In the first step of the invisible problem-solving loop, the three-level FDD approach is used to explore the data to uncover deviations from normality that may provide useful information for the development of a data-driven solution.

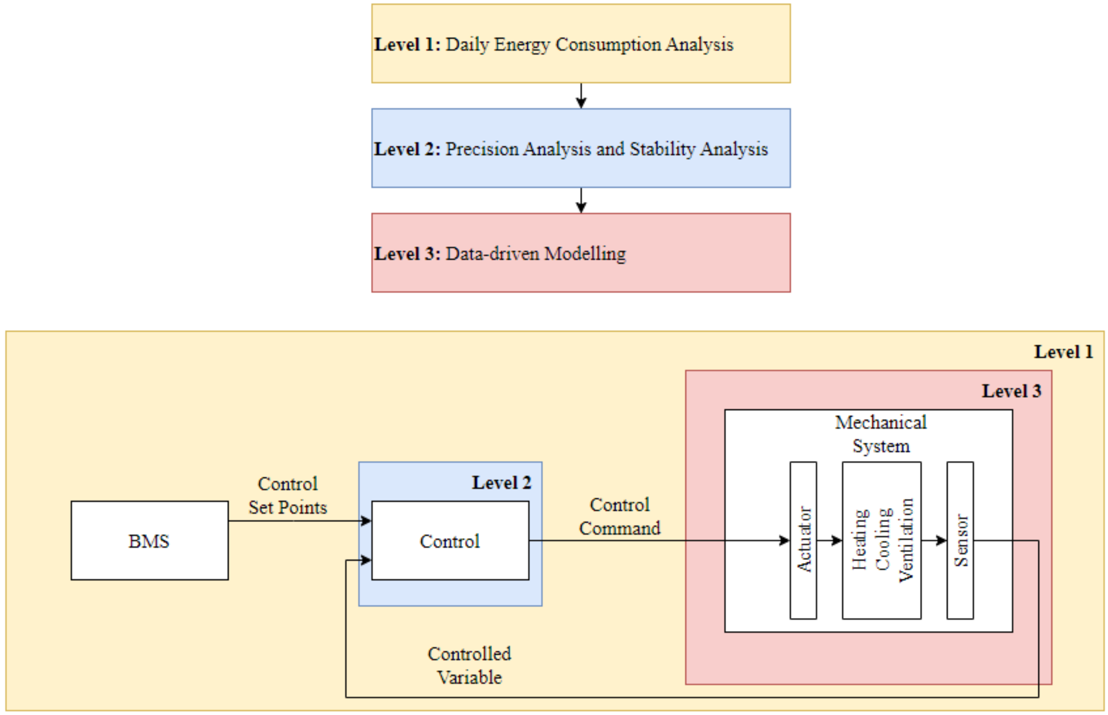


Figure 5-23 Three-level approach to invisible problem-solving, adapted from Gao et al. [189]

5.3.2 Data Exploration

Figure 5-24 Data Exploration phase of the IDAIC framework

The first level of the data mining-based fault detection approach proposed by Gao et al. [189] is an analysis of daily energy consumption. As discussed in subsection 5.3.1, in the absence of accurate flow rate readings, the analysis must be simplified. The thermal energy consumed may be estimated using the temperature difference across the coil (ΔT), and by assuming that the design mass flow rate through the AHU (m) is delivered at 20 kg/h, a specific heat capacity of air (c_p) of 1.005 kJ/kg K and boiler efficiency of 80% or chiller COP of 3 (Eff) for heating and cooling respectively. The daily thermal energy consumed is then defined by Eqn (1):

$$Daily\ Heating/Cooling\ Energy\ Consumed = 24 \times \frac{m \times\ c_p \times \Delta T}{Eff} \qquad \text{Eqn (1)}$$

The lack of labels in this dataset increases the difficulty of uncovering valuable insight as it is difficult to determine normal operation from faulty operation. However, the understanding obtained from the visible problem-solving phases may be used to develop labels that are known to affect the energy consumed by this unit. This would help to address the operational gap. For example, the COVID_Operation resulted in a significant amount of time spent in mechanical heating and mechanical cooling modes. The impact of this abnormal control strategy may be compared to normal operation through comparison with the average daily outside air temperature. Given the availability issues for the fan energy meters identified in the data assessment phase, the last two years of the dataset are selected for comprehensive analysis of the energy-consuming components in the AHU. Firstly, the analysis of the heating coil uncovered that at lower outside air temperatures, more heat is consumed in full fresh air conditions than in normal operation as shown in Figure 5-25.

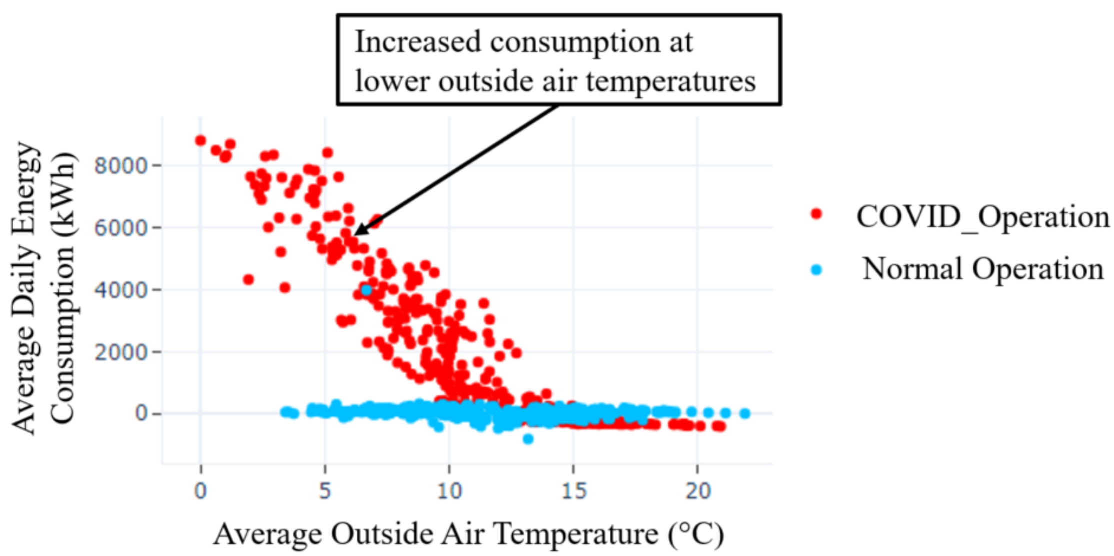


Figure 5-25 Level 1: Average daily energy consumption vs average daily outside air temperature for the heating coil

Secondly, in Figure 5-26, COVID_Operation resulted in more cooling energy being consumed as the outside air dampers could not reduce the fraction of hot outside air entering the AHU.

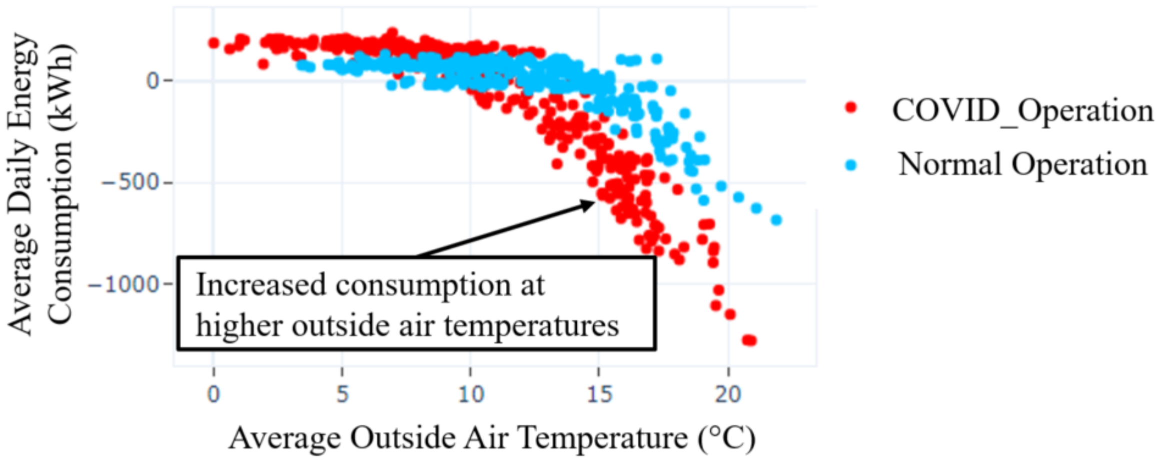


Figure 5-26 Level 1: Average daily energy consumption vs average daily outside air temperature for the cooling coil

Little value is obtained from the daily fan energy consumption subplots as the consumption is not influenced by the outside air temperature as shown in Figure 5-27. Given the fixed nature of the fan consumption and lack of airflow data, detecting signs of an invisible fault, such as decreased efficiency, becomes difficult and time-consuming.

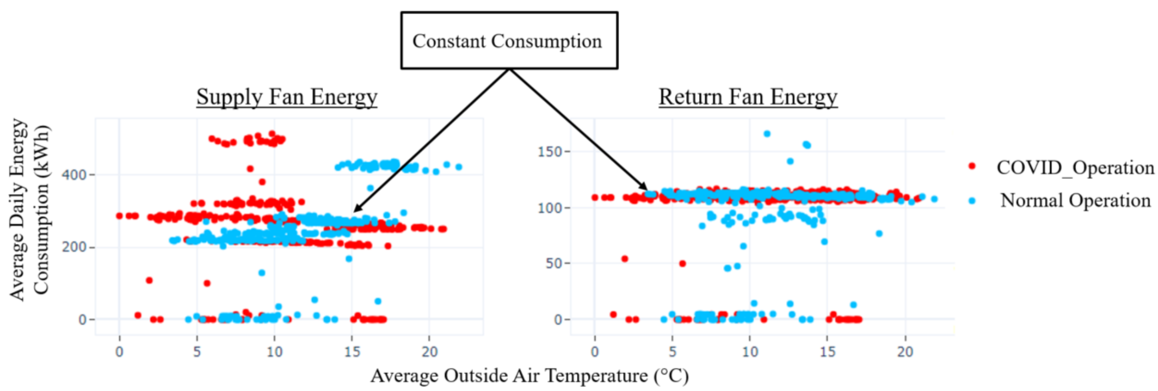


Figure 5-27 Level 1: Average daily energy consumption vs average daily outside air temperature for the supply and return air fans

Figure 5-28 highlights an example of an issue that may cause industrial practitioners to lose faith in data-driven approaches. The bias fault detected in the data assessment phase between the zone supply air temperature sensor and the supply air temperature sensor in zone 2 has led to an inaccurate estimation of the heating energy consumed by the air passing the reheat coil. This further highlights the need for domain integration in practical data-driven analysis to ensure the analysis provides value and to avoid inaccurate results.

Meanwhile, the COVID_Operation appears to have had no impact on the heating consumption in zone 1.

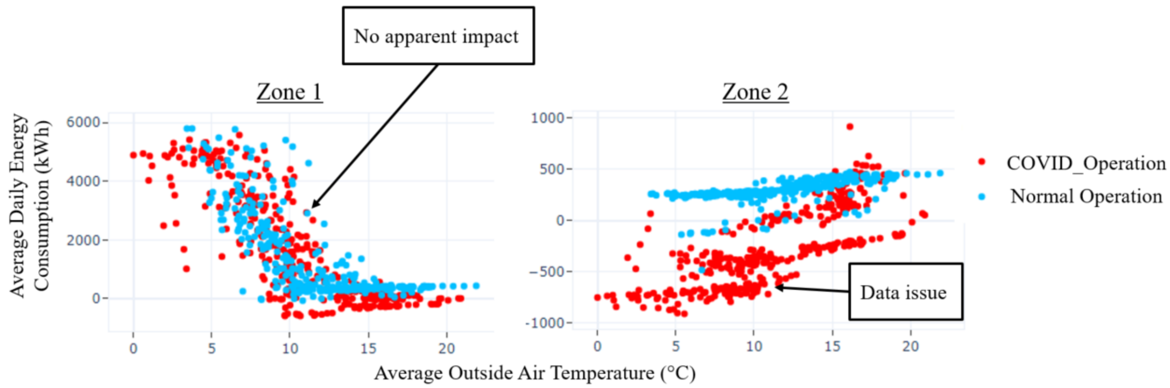


Figure 5-28 Level 1: Average daily energy consumption vs average daily outside air temperature for the reheat coils

In level 2 of Figure 5-23, the precision and stability of the system are analysed for control performance evaluation. While the zone air temperature control has not been fully investigated in the visible problem-solving space due to the complexity of controlling two different zones, Figure 5-29 and Figure 5-30 investigate the control using a data-driven approach.

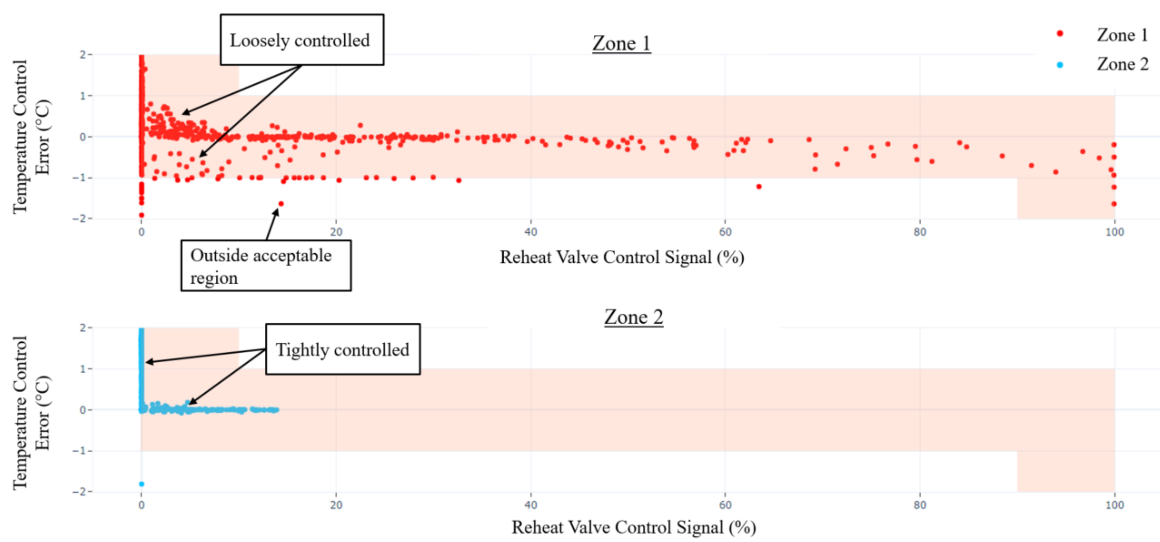


Figure 5-29 Level 2: Precision Analysis of average zone temperature control error vs average reheat valve control signal for zone 1 and zone 2

In Figure 5-29 both zone 1 and zone 2 temperatures appear to be well controlled in the range of ±2°C from their respective set points. However, negative outliers exist beyond this range for both zones when the average valve control signal is fully closed, which suggests that the control strategy should be reviewed to ensure that sufficient heating is provided to the zones. While most of the remaining data points lie within the accepted shaded region as defined in the analysis by Gao et al. [189], it appears that zone 2 is more tightly controlled as the average control error is generally much closer to 0°C than zone 1. Zone 1 appears to require more heating based on the average reheat valve control signal which ranges from fully closed to fully open, while the average control signal for zone 2 does not exceed 20% open. This is expected to be the case considering that the zone 1 set point is 1°C higher than zone 2. In Figure 5-30 the daily cumulated absolute temperature control error appears to show a general trend of tight control. However, the zone 2 temperature suddenly began to show significantly greater deviations from the set point than before this point. This began just before the known control logic review in December 2021. While the zone temperature set point may have been changed, it is also possible that the deadband for this zone was increased. This may be an issue for this zone if the equipment or product requires tight temperature control. Similarly, the cumulative zone 1 error appears to increase after the shutdown in May 2022. Both of these findings indicate that the zone temperature control strategy has changed either by changing the zone temperature set point or increasing the deadband range. A follow-up with the BMS vendor is needed to determine the actions that took place so that the thresholds for the zone temperature control are appropriately updated.

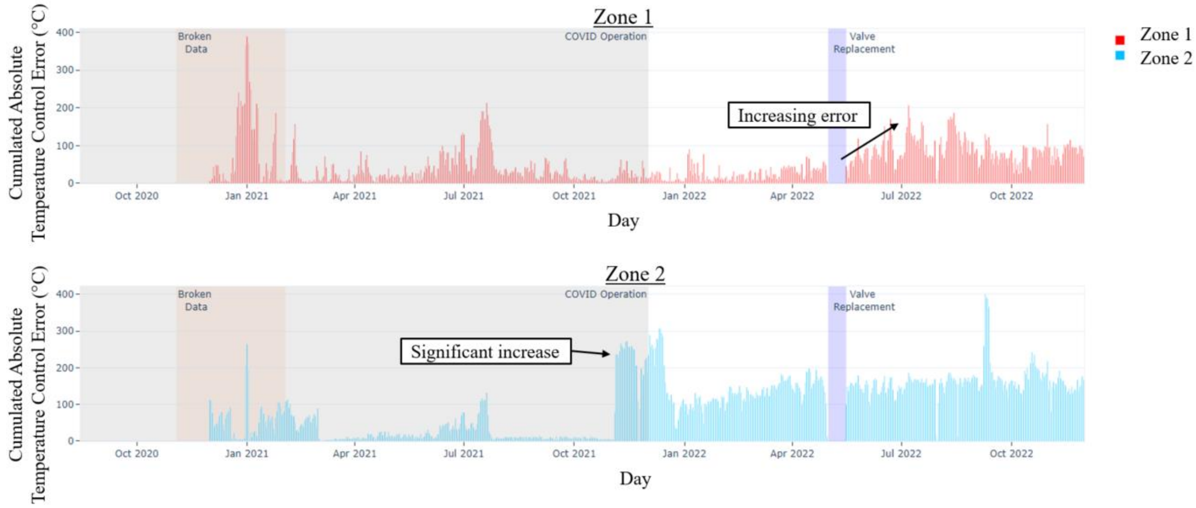


Figure 5-30 Level 2: Stability Analysis using the cumulated absolute change of zone temperature control error (°C) for zone 1 and zone 2

In level 3, the performance of each mechanical component is analysed. Heating power, cooling power, supply airflow and return airflow are proposed for modelling by Gao et al. [189]. As the dataset in this case study is not labelled, more exploration is required to determine the "near perfect" operation from the "faulty" operation, also known as the operational gap. Figure 5-31 visualises the relationship between features that are known to relate to the performance of the outside air damper. As identified as a shortcoming of supervised approaches in Chapter 2, the COVID_Operation greatly restricts the range of conditions to train a data-driven algorithm for detecting invisible damper performance degradation, given that the outside air damper is set to the fully open position. This further shows the need for integrating domain expertise to provide background and context to the data before algorithm application.

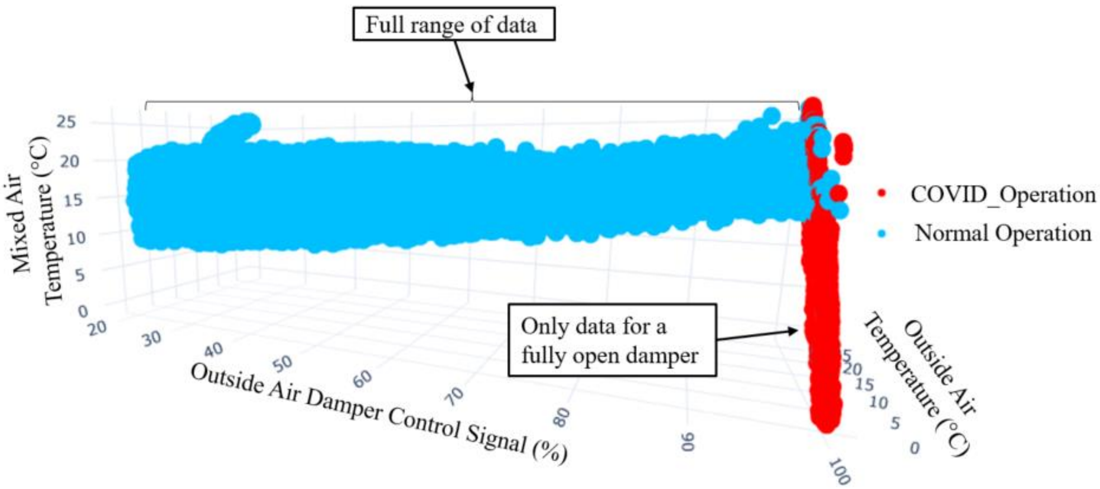


Figure 5-31 Level 3: Mechanical Component Analysis of the outside air damper

The remaining invisible faults from Table 5-3 include leakage of valves and fouling of coils. Given the lack of water temperature and flow data, the air-side temperature sensors and control signals are used to analyse these faults in Figure 5-32 and Figure 5-33. The last two full months of data, October and November 2022, are used to explore the most recent relationship between the features. In Figure 5-32, the performance of the heating coil is analysed by visually inspecting the relationship between the hot water valve control signal, the air temperature before the heating coil (MaTemp) and the temperature difference across the coil. Given that the hot water valve control signal is fully closed during this period, a negligible temperature difference should be seen across the coil. However, a large cluster appears to develop that indicates the temperature difference is between -2°C and +2°C. This makes it difficult to detect subtle leakages and would also impact the ability to detect fouling if the valve was opened. Similarly in Figure 5-33, the performance of the cooling coil valve is analysed. When the valve control signal is fully closed in November 2022, a similar cluster and associated issues as identified in the heating coil are observed. However, the chilled water valve opened in October 2022, which enables a rudimentary analysis of the cooling performance of this AHU. While these three-dimensional plots enable an analysis of a subset of invisible faults, many occurrences of temperature drops across the heating coil in Figure 5-32 suggest that further commissioning work is required before trust may be placed in a fully data-driven solution.

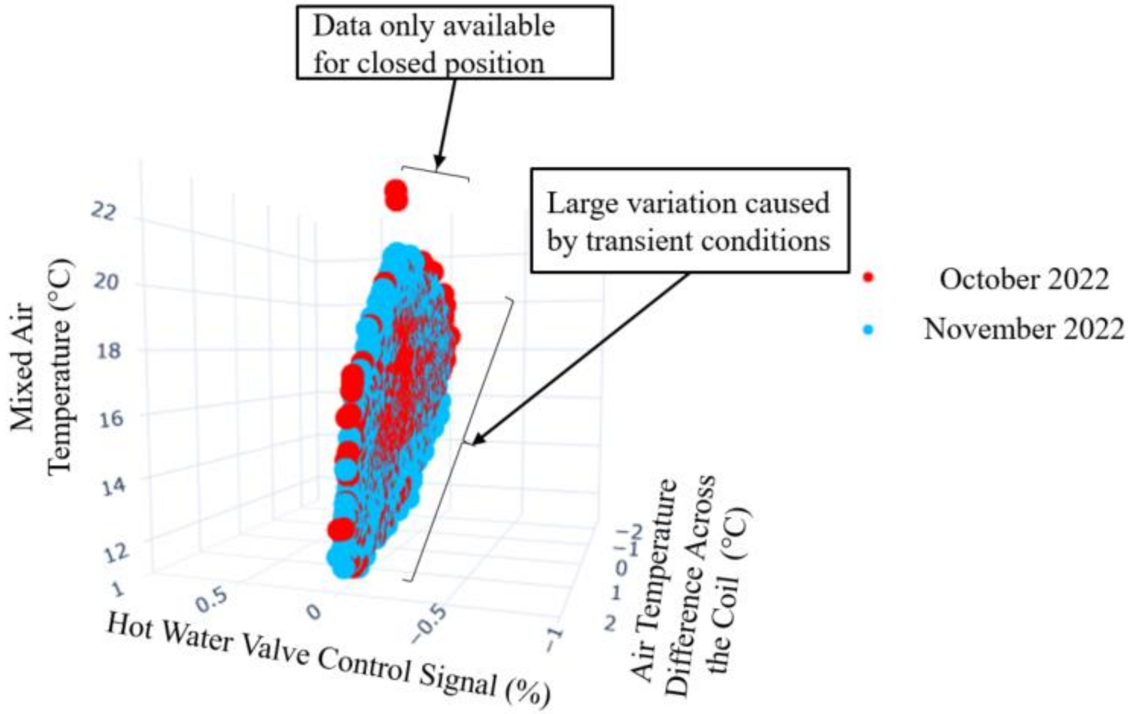


Figure 5-32 Level 3: Mechanical component analysis of heating coil

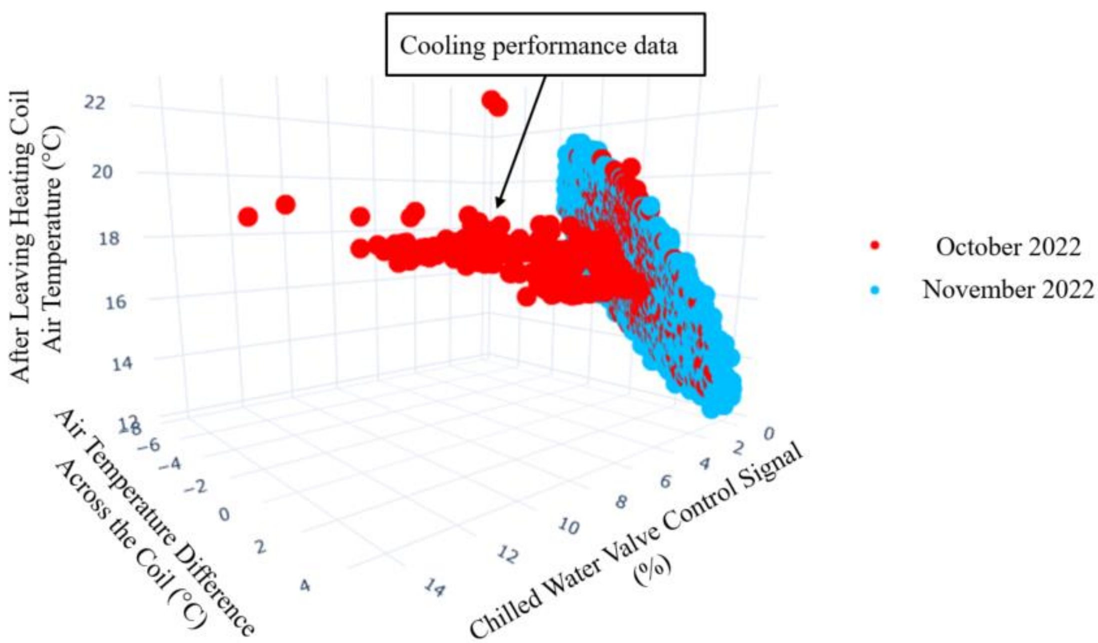


Figure 5-33 Level 3: Mechanical component analysis of cooling coil

5.3.3 Algorithm Selection & Results Exploration

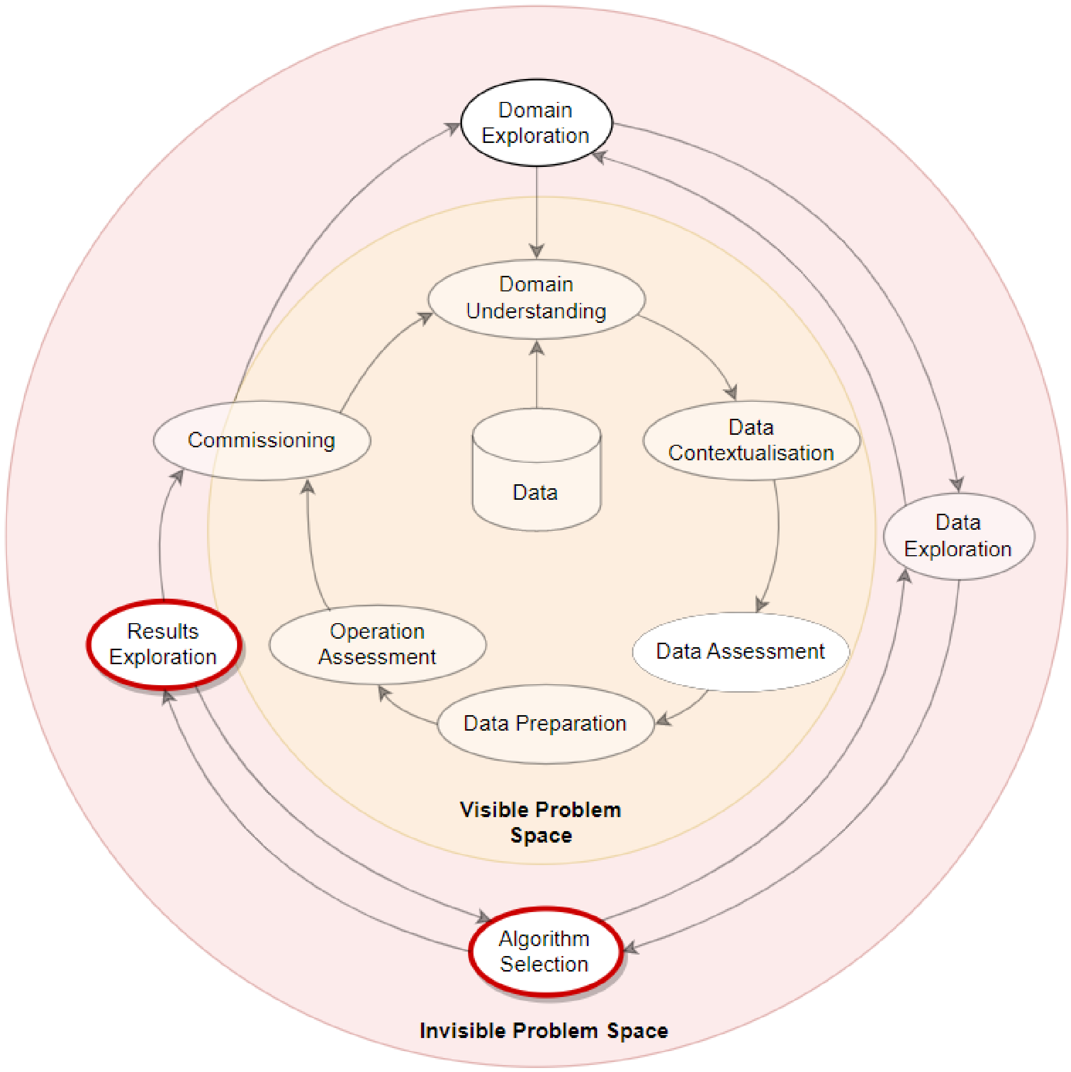


Figure 5-34 Algorithm Selection and Results Exploration phases of the IDAIC framework

Given that the case study site only partially addressed the identified visible problems, full implementation of the invisible problem-solving loop was not conducted on account of establishing a misleading fault-free baseline. As in Table 2-9, both supervised and unsupervised techniques require high-quality data which is not currently the case. The analysis of the data should therefore remain in the data exploration phase whereby domain expertise may be leveraged to ensure the results are correctly interpreted and highly explainable. If the analysis becomes black-box in nature with inadequate data quality, the results may be inaccurate which could lead to a lack of trust in proactive approaches in the practical setting. To address this issue, the faults identified in the data assessment and operation assessment phases must be fully resolved in the commissioning phase. Once the AHU is operating in a visible fault-free state, the focus of the analysis may turn towards data-driven invisible problem-solving.

The selection of an appropriate algorithm for data-driven FDD is key given the generalisation issue referenced in the literature. As discussed in Chapter 2, recent work as of 2023, has proposed diagrams that aid the selection of an appropriate algorithm [105]. This approach appears to be a suitable means of guiding the selection of an algorithm. Given, the unlabelled nature of the dataset in its current form, the diagram developed by Matetic et al [105] would correctly suggest that only unsupervised learning is applicable. However, given that domain knowledge has been systematically integrated through the implementation of the IDAIC framework, some labels may be applied to the dataset. For example, the APAR ruleset labels the dataset for particular conditions that have been met, along with mode identification. These labels could then be used to further inform the selection of an appropriate algorithm. This concept is expanded further in the upcoming tool deployment chapter to show how a hybrid approach may be taken to algorithm selection.

5.4 Results, Impact & Discussion

The effectiveness of the IDAIC framework is outlined in Table 5-11 in relation to its ability to address the data and operational gaps, the second and third research questions respectively. Firstly, the challenges relating to the data in this case study have been largely addressed through the implementation of the IDAIC framework. The availability of the data has increased from 2.5 to 5 years due to the integration of offline (Cylon) and online (Tridium) datasets. The quality of the dataset has been quantitatively measured using a combination of statistical methods and domain expertise to identify common sensor faults. While both the utility and understanding of the data are significantly improved through the 17 data preparation tasks and labelling of the dataset for both steady-state and COVID_Operation. Secondly, the impact of the implementation of the IDAIC framework on the operational gap is compared in relation to the ten desirable characteristics of FDD methods as proposed by Venkatasubramanian et al. [191] and utilised on AHUs by Yu et al. [139]. The main improvement is in the robustness of the FDD as the applicability of the expert rules have increased from 3 years to 5 years for this dataset by extending the APAR rules for COVID_Operation. The ability to inspect subtle deviations from normality in the invisible problem-solving space also improves novelty identifiability. This is achieved through three levels of practical analysis that aid the identification of faults through anomalous daily energy consumption, unstable or imprecise control, and abnormal mechanical component performance. The data issues in this case study would lead to high classification error without IDAIC, however, the data assessment phase reduces the likelihood of performing unnecessary maintenance by correctly identifying data issues early in the analysis. As IDAIC is a domain-integrated framework, the FDD is highly adaptable and explainable. However, as the analysis progresses to data-driven methods with capabilities for multiple fault identification over time, storage and computation challenges may potentially increase.

These findings along with other contributions are discussed in further detail from the perspective of each problem-solving activity in the following subsections.

Table 5-11 Comparison between the cases with and without the IDAIC framework

Category	Challenge	Without IDAIC	With IDAIC
Data	Data Availability	Limited to 2.5 years	5+ years
	Data Quality	Unknown	Specific sensor faults identified
	Data Utility	Unknown	17 data preparation tasks were identified and completed to improve the usefulness of the data
	Data Understanding	Unknown	Identification of approximately 47% transient state and approximately 17 months of COVID_Operation which aided the applicability of APAR through modifications
Operation	Quick detection and diagnosis	Yes	Yes
	Isoability	Yes	Yes
	Robustness	Low	High
	Novelty identifiability	Low	High
	Classification error	High	Low
	Adaptability	Low	High
	Explanation facility	Medium	High
	Modelling requirement	Low	Low
	Storage and computation	Low	Medium
	Multiple fault identifiability	Low	High

5.4.1 Visible Problem Space

The activities carried out in the visible problem space resulted in key findings relating to data challenges and knowledge-based FDD.

5.4.1.1 Data Challenges

The literature suggests that data issues are a significant barrier to proactive approaches in the real world and the case study confirmed this. Before the implementation of IDAIC, the data may have been considered unusable given the siloed nature of the heterogeneous data sources. However, through strong communication with the onsite facilities team, the datasets were integrated to improve the overall availability of the data from 42.06% for the cloud-based data, and 63.34% for the offline data logs, to an overall availability of 86.04% for the 66 months available. However, some key features were missing from the dataset which requires a digitalisation plan to be developed to acquire this key data. From a data quality perspective, the identification of the numerous challenges in Table 5-8 would not have occurred without the application of the data assessment phase and the development of a novel data assessment decision tree. Among the most severe faults include the drift and bias fault in the outside air temperature sensor in Figure 5-13 and the spike faults in the zone air temperature sensor in Figure 5-12. Furthermore, the utility of the dataset is greatly improved for the 16 months of COVID_Operation by investigating its background. The analysis of the summary statistics of the similar sensors in Table 5-7 led to the identification of two main background issues; the increasing occurrence of transient conditions as in Figure 5-14, and COVID_Operation as shown in Figure 5-19. Transient conditions are difficult to analyse given the possibility of lost information between the 15-min samples, while the manually overridden outside air damper control signal during COVID_Operation causes the APAR ruleset to become non-applicable. Therefore, implementing IDAIC created a real-world 5-year dataset that could be utilised by identifying data issues in a structured manner leveraging domain expertise. A summary of the contributions and associated impact relating to the data issues are shown in Table 5-12

Table 5-12 Novel solutions and contributions to the gaps identified in the literature

Problem-Solving Space	Gap	Novel Contribution	Impact
Visible	Lack of a process to determine, and improve, real AHU dataset usefulness	Developed a novel AHU data assessment decision tree to detect data issues	Broken - ≈ 23% increase in data comprehensiveness Bad-quality - 17 issues identified and mitigated Background – >20,000 hours of transient state operation detected
	Lack of pilot analysis on real-world case studies to tackle practical issues and aid reproducibility	Extended APAR ruleset for application in the novel modes of operation created in response to the COVID-19 pandemic	Reduced 16 months of false alarms that would lead to unnecessary maintenance
	There is a lack of an indicator to determine when the data-driven analysis is feasible	Developed semantic consistency rules to measure variation in the data	Three quantitative metrics were developed to measure progress to data-driven analysis; Number of visible faults The proportion of time spent in a transient state Average variability of the air temperature sensors
Invisible	Lack of domain-guided data-driven analysis	Practical amendments for real-world data-driven analysis using proxy indicators	Identification of missed opportunity in control logic review of approximately €60,000 annualised savings

5.4.1.2 Knowledge-based FDD Challenges

Once these issues were identified, the necessary data preparations and FDD amendments were required to apply knowledge-based FDD for visible problem-solving. If left unattended, the data issues would result in many falsely triggered faults and time-consuming fault diagnoses. For example, the drift fault detected in the outside air temperature sensor would incorrectly trigger rule 10 in Mode 3 as it would exceed the threshold for the absolute difference with the mixed air temperature sensor. House et al. [141] suggest four separate explanations for this triggered rule which leads to unnecessary costly maintenance activity taking place. In the IDAIC framework, the occurrence of unnecessary maintenance is reduced by early detection of sensor faults. These sensor issues may then be addressed to ensure that they align more closely to the true physical property they measure, such as the replacement of a faulty sensor with values from another sensor in the case of the outside air temperature sensor.

Similarly, the effectiveness of the FDD method has been increased by amending the ruleset based on the data assessment and domain expert input. In this case study, the application of the APAR rules needed to be amended to account for controlling to the zone air temperature set point, the transient state conditions, and COVID_Operation. Firstly, the supply air temperature and the associated set point were replaced in each rule with the zone air temperature and associated set point. An "OR- condition" is included given the presence of two zones. Secondly, APAR can deal with transient conditions by obtaining an exponential weighted moving average of the data [141]. However, given the 15-min interval granularity of the data, the information captured during transient states is deemed to be insufficient for visible problem-solving. Therefore, transient states are considered as a control fault and left for further investigation in the invisible problem space given the lack of visibility into the data. Thirdly, the COVID-19 operation led to the identification of novel modes of operation for this type of AHU. The APAR ruleset was subsequently expanded to detect this operation by identifying a continuously fully open outside air damper for a period of one week. The modes are then classified as free cooling, mechanical cooling and mechanical heating based on the hot water and chilled water valve control signals. Therefore, without IDAIC, FDD could not be applied correctly. To enable visible problem-solving, a structured integration of domain expertise is used to make informed decisions regarding data preparation and FDD amendments.

5.4.2 Invisible Problem Space

To make the transition to invisible problem-solving, the issues identified in the visible problem-solving space must be fixed. As in Table 5-10, not all of the issues identified in this case study were resolved, but some have been. This means that the invisible problem-solving analysis may progress using domain knowledge for interpretation.

5.4.2.1 Incomplete Maintenance

The ad hoc nature of the maintenance activity on this unit meant that little documentation regarding works completed is available. Furthermore, the effectiveness of this activity has previously been unknown before the IDAIC implementation. During this case study, the analysis of the data led to three main maintenance activities taking place; sensor re-calibration, control review, and valve replacements. The sensor re-calibration and control review occurred as a result of the visible problem-solving activity. The effect of this work may be seen in the example in Figure 5-35. The sensor calibration appears to show an improvement as the mean values of SC_3 appear to be closer to 0, indicating the mixed air temperature sensor and air leaving the heating coil temperature align more closely. Also, the effect of the hunting dampers may be seen by the generally larger interquartile range for transient damper control in orange than for steady-state damper control in blue. A significant improvement is also seen regarding the total number of visible faults in Figure 5-19 after the control logic review in December 2021. The total number of faults reduced from over 5000 in November 2021 to a total of 91 faults in January 2022 and 85 faults in February 2022. However, an increase occurs in March 2022 as a result of the increased mode transitions discussed in subsection 5.2.5, and the change observed in zone temperature control regarding set points and deadband as discussed in subsection 5.3.2. While this is possibly an example of a common shortcoming with rule-based systems, the need for continuous upgrades [70], the detection of a fault in the visible problem-solving space provides key information for invisible problem-solving. The zone temperature control precision and stability analysis, in Figure 5-29 and Figure 5-30 respectively, could lead to the development of an inaccurate data-driven solution if the assumed set points are incorrect. This further highlights the need for comprehensive maintenance in the commissioning phase, which is sufficiently documented, before developing a data-driven solution.

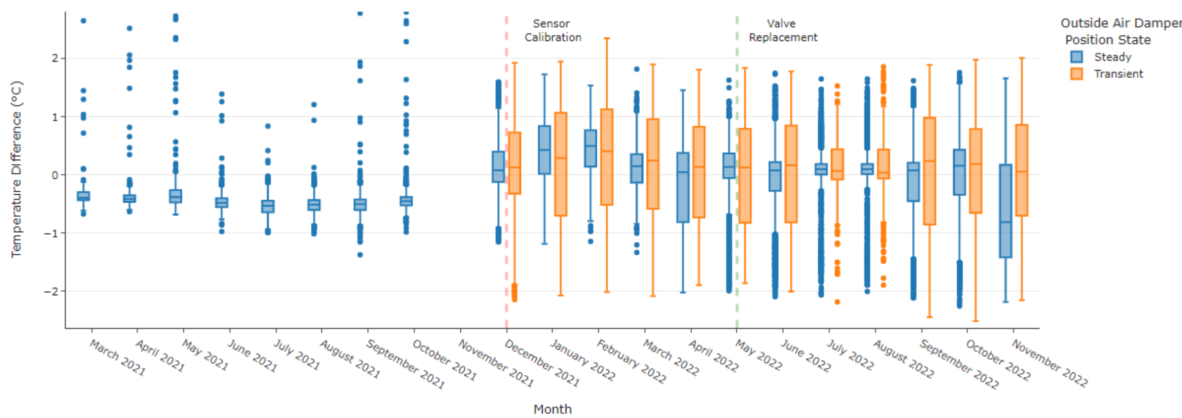


Figure 5-35 Monthly boxplot of the air temperature difference across the heating coil when the valve is closed

5.4.2.2 Data-Driven FDD Challenges and Practical Adaptations to Enable Analysis

The challenges relating to data-driven FDD from the literature were presented in Table 2-9. The findings from this case study further highlight and contextualise these challenges. For example, the lack of labelled data caused difficulty in determining the occurrence of COVID_Operation. Furthermore, the failure to rectify the outside air temperature sensor error and transient state control leads to an unbalanced faulty dataset which causes further difficulty in labelling component faults for the training of a data-driven method. Similarly, it is also difficult and time consuming to understand the relationships, clusters, and anomalies discovered in the data exploration phase and to then determine which abnormalities should lead to maintenance actions. While in relation to investigating specific faults, significant effort is required to develop useful analysis when key features are missing. However, the need for high-quality data in both supervised and unsupervised approaches is perhaps the most significant challenge facing the adoption of data-driven approaches. The implementation of the IDAIC framework in this chapter focused on the identification and mitigation of data issues. Given that most of the data is contained in a siloed system, the data has not been previously analysed regarding the relationships between features, cluster analysis or anomaly detection for example. The IDAIC framework now enables the partial analysis of energy consumption and component degradation through the practical adaptation of data-driven methods. Given the significant proportion of time spent in transient states, a data-driven data mining method proposed by Gao et al. [189] is applied given the use of daily averages. While the data issues limited the accuracy of the analysis, assumptions and estimations made with domain expertise enabled progression towards data-driven analysis. The thermal energy consumed by the unit could be estimated and used to detect anomalies or measure the impact of operational decisions. For example, the unit remained operating in COVID_Operation for a significant proportion of time after the decision was made to return all the units to normal operation. A key reason why the missed control strategy change went unnoticed is that this data was not visible before the implementation of IDAIC.

Utilising the data for the heating energy in the main unit as visualised in Figure 5-25, the annualised thermal heating energy consumed in the period from December 2020 to December 2021 is estimated to be 742,348.75 kWh using Eqn (1). Given that the heating coil valve control signal remained closed from December 2021 to December 2022, this may be considered excess heating and at an assumed cost of gas of 8.59 c/kWh [192], an annual excess heating cost of €63,767.75. The data-driven analysis has also led to increased visibility regarding component degradation. In Figure 5-31 and Figure 5-33, a visual representation of the relationship between the mechanical components in this AHU is developed to enable visual detection of clusters and anomalies that may be indicative of degradation. Given the maintenance actions that remain, a black-box algorithm is not developed as high data quality is required before trusting the results of an approach that cannot be explained. The summary of the high-level findings of this case study is presented in Table 5-13.

Table 5-13 Recommended Actions

Problem-Solving Space	Phase	Barrier	Future Work
Visible	Data Assessment	Missing Data	Conduct a digitalisation plan to acquire more key features
		Poor Data Quality	Some internal air temperature sensors re-calibrated. Further recalibration or replacement of the outside air temperature sensor is needed
		Transient Control	Control strategy reviewed. Improvement seen but further review needed to address hunting control signals
	Operation Assessment	Many triggered faults	Steadier control is required to analyse further
Invisible	Commissioning	Incomplete actions	Further work is required to transition to invisible problem-solving
	Data Exploration	High variability in the data	Analysis remains highly interpretable until the necessary maintenance actions take place

5.4.3 Potential for Proactive Maintenance

The final two phases of the framework were not applied due to the presence of unresolved visible problems. However, preliminary results indicate that a proactive maintenance approach offers several benefits. It incorporates historical data for scheduling maintenance actions, helps achieve emissions targets, and promotes improved system operation. APAR predominantly only looks at the value obtained for each sensor at a particular instance in time or over a short window such as an hour for rule 28. Therefore, it performs poorly in a proactive invisible problem-solving approach that is focused on detecting the degradation pattern that is indicative of faults in mechanical equipment over time. If the necessary maintenance activity occurs to enable this transition, the added benefit of analysing the degradation over time would improve the scheduling of maintenance actions so that they take place at the appropriate time. Secondly, according to the IEA, a third of emissions reductions must come from improvements in energy efficiency [24]. A proactive approach would ensure that inefficiencies are avoided ahead of time. Thus contributing to climate targets. Lastly, better monitoring leads to better operation. For AHUs, this leads to better thermal conditions and IAQ, which is known to have a positive impact on employee cognitive function [193].

The potential to reproduce similar analysis, as outlined in this chapter, is strongly dependent on the availability of domain expertise. As noted by Karniadakis et al. [194] in relation to physics-informed ML, the more data the less need for physics, and the less data the more need for physics. Concerning industrial data analysis, the quantity, quality and understanding of the data must be accounted for before the application of FDD. In the IDAIC framework, the data assessment phase quantitatively measures each of these features of the dataset. Therefore in the context of real-world applications using the IDAIC framework it may be stated that the better the overall data assessment score the less need for domain expertise and vice versa. It is difficult to quantitatively measure exactly how much domain knowledge would be required in other applications, but the data assessment phase gives an early indication of the data gaps that require a physical understanding of the equipment to fill.

5.5 Conclusion

In this chapter, the IDAIC framework is applied to an AHU in a manufacturing facility to develop a proactive maintenance strategy. While a black-box data-driven solution should not yet be applied given the presence of data issues, the case study led to the following impactful outcomes and novel contributions. Firstly, a roadmap for proactive maintenance on this AHU has been developed that has led to positive changes regarding the operation, energy consumption and data quality that may be replicated on other AHUs. Secondly, the lack of real-world datasets in the literature has been addressed by providing the data used in this case study which spans five years. Thirdly, a novel data assessment decision tree has been developed to detect sensor faults in AHUs using domain knowledge to derive similar sensors. Finally, extensions to the popular APAR ruleset have been developed to account for the novel mode of operation of this AHU during the COVID-19 pandemic.

The main finding from this research is that significant integration of domain expertise is required to contextualise and supplement insufficient real-world datasets. The quantity and variety of practical issues are likely to discourage early adopters of proactive strategies without the aid of a structured and flexible framework. The IDAIC framework proved to successfully identify and mitigate many of the issues in this case study but actions that need to be physically carried out, such as control logic review and sensor recalibration, remain a barrier. Therefore, the transition to proactive maintenance depends on the willingness of key stakeholders in the industrial facility to provide the resources to maintain a high-quality dataset to train data-driven solutions. Nevertheless, the research carried out in this case study has resulted in the development of an analysis that assists decision-makers in making well-informed maintenance-related decisions. For example, one of the recommendations was to revert to the use of dynamically controlled air dampers, which is estimated to yield annual savings of €60,000.

In future works the applicability of the framework to other AHUs would be improved by further refining and automating the data assessment phase to detect data issues and provide more accurate data imputation techniques in the data preparation phase. If the case study site decides to implement the necessary maintenance actions to resolve the issues identified in this case study, then a data-driven approach may be applicable and deployed through a visual-based tool to aid decision-makers. Lastly, the framework could be applied to another type of industrial mechanical equipment to test its applicability to another system.

6 Chapter 6 – Tool Deployment

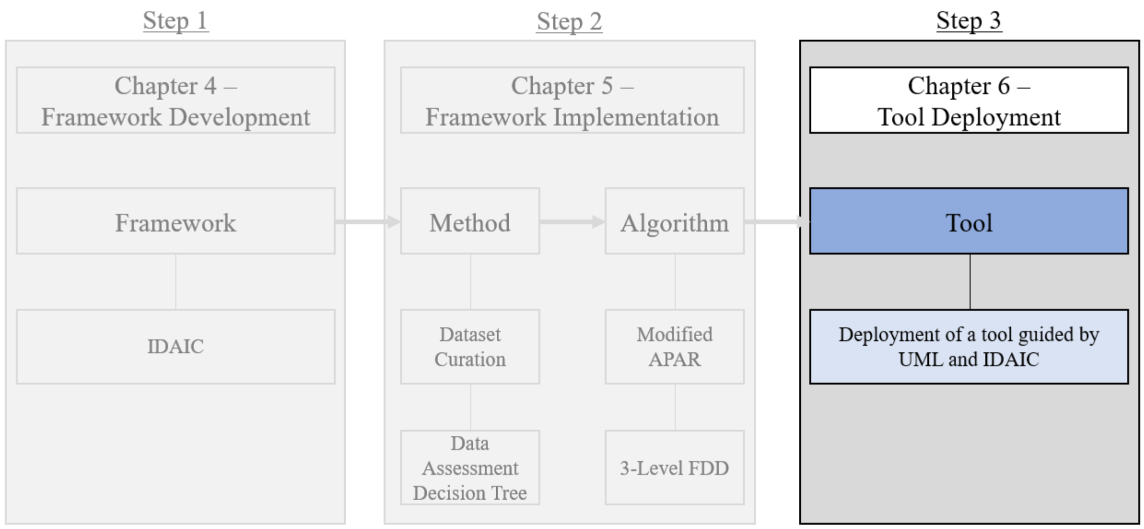


Figure 6-1 Step 3 of the research methodology

6.1 Introduction

The objective of this chapter is to document the development of a tool underpinned by the IDAIC framework, at a large manufacturing company. The tool development process will help to uncover the practical considerations that contribute to the practice-theory gap in the industrial setting, as well as determine the ability of the IDAIC framework to address these real-world factors, i.e. to enable reproducibility of the analysis. Therefore this would address the fourth, and final, research question *"What practical considerations should be taken into account to enable reproducibility in the development of a tool for the purposes of proactively maintaining an AHU in the industrial setting?"*

The tool was developed in collaboration with partners on the DENiM EU project [195]. DENiM stands for digital intelligence for collaborative energy management in manufacturing and is an EU project coordinated by MTU with multiple partners engaged throughout Europe. Among these partners, four serve as pilot sites that function as testbeds for the practical development and application of digital technologies in real-world scenarios. One of the main aims of this research consortium is to equip a pilot site with a tool that enables continuous monitoring and optimisation of energy usage, along with fault detection capabilities. The most energy-intensive assets on the pilot site are selected for inclusion, thus AHUs are within the project scope. Therefore, the DENiM project is a suitable means of deploying the research outlined in this thesis on a real-world site. The contributions of this research and the contribution of the DENiM group are clearly outlined in the following subsections. The requirements of the tool and the logic to achieve these requirements is the focus of this research, while the delivery of this research in terms of the software engineering-based tools are provided by the DENiM partners.

The deployment of this tool is critical to showing the effectiveness of this research as it enables continuous monitoring of an AHU and quantification of the efficiency savings identified by implementing this framework. The research thus far has completed the first two steps of the research methodology, progressing from the framework level to the algorithm level. In Chapter 4, the IDAIC framework development process was documented using process models for visible and invisible problem-solving to guide the transition to proactive maintenance which addressed the knowledge gap. In Chapter 5, the IDAIC framework implementation process was documented using methods and algorithms which addressed the data and operational gaps. These research outputs now enable the final step, the development of a tool to address the practice-theory gap at the tool level.

While the combination of the knowledge, data, and operational gaps undoubtedly contributes to the lack of deployed tools in the industrial setting, other practical considerations may arise through the integration of a tool into the maintenance process. Some of these considerations may be elicited through the actual deployment of a tool in a case study site. Then, by using the IDAIC framework to guide the development of the tool in an industrial facility, the technical barriers that contribute to reproducibility and practicality issues may be tackled. Therefore, in subsection 6.2 the Unified Modelling Language (UML) is used in collaboration with the IDAIC framework to diagrammatically document the design of the tool. In subsection 6.3, the technical implementation of the tool is documented. The practical considerations that were encountered are then discussed in subsection 6.4, along with the ability of the IDAIC framework to address these issues. A critical appraisal is then provided in subsection 6.5.

6.2 Tool Design Using UML

The IDAIC framework provides a guide to the transition to proactively maintaining AHUs in the industrial setting. However, to translate the implementation described in Chapter 5 to a deployed tool, a structured means of developing a software tool must be followed. The Unified Modelling Language (UML) is the de facto industrial standard for the development of software systems [196]. It is a means of developing a diagrammatic and intuitively understandable contract between a client (user) and a supplier (computer scientist) [196]. Similarly to the IDAIC framework, it aims to bridge a knowledge gap between users that do not understand the intricacies of software development and system builders who do not understand the specific domain. As outlined in Martin Fowlers' UML Distilled [197], UML is called a modelling language, not a method. Fowler [197] further explains that a method generally consists of both a mainly graphical modelling language that expresses designs, and a process on the steps to take when undertaking a design. Given that UML is designed for facilitating communication on software development, it may be integrated with the IDAIC framework to provide a tangible means of relaying the necessary activities in the tool deployment process on an "as needs" basis. Therefore, the combination of UML and the IDAIC framework would provide a common graphical language, along with a structured, yet flexible, approach to tool development.

The changing structure in UML can cause difficulty in selecting the appropriate diagrams to use. For example, the main classification of diagrams in the first edition of the UML reference manual is defined as structural, dynamic and model management [198]. There are multiple diagrams then outlined. For example, the structural diagrams include the class diagram, use case diagram, component diagram and deployment diagram [198]. Then, in the second edition of the UML reference manual, the main diagram classifications are updated to structural, dynamic, physical and model management [199]. The deployment diagram is then classified as physical in the second edition [199]. In later years, UML updated to version 2.0 where there are thirteen diagrams divided into three categories; structure, behaviour, and interaction [200]. These appear to have remained as the preferred classification types in the latest version (UML 2.5) [201].

In a systematic literature review on the use of UML diagrams, Koç et al. [202] found that class diagrams, activity diagrams and use case diagrams are the most commonly used in publications. Taking into account the prevalence of these diagrams in literature, it indicates a tendency towards their recognition as preferred graphical representations for depicting key aspects of tool deployment. Therefore they have been selected as the means of documenting the tool development process. Additionally, given the deployment nature of this chapter, the deployment diagram is also included. The interaction between the IDAIC framework phases and each of these four UML diagrams is depicted in Figure 6-2.

The purpose of the UML integration is to communicate the different aspects of the tool development process. As shown in Figure 6-2, the selected UML diagrams are aligned to the relevant IDAIC phases. As IDAIC is based on the CRISP-DM process model, the phases do not necessarily need to be completed in sequential order, which suggests that the diagrams do not necessarily have to be in order either. However, some linkages do exist between the diagrams, which may be logically presented as follows; firstly, the data is contextualised in the class diagram. Secondly, the requirements of the tool are outlined in the use case diagrams. This naturally leads to the activities to be completed to achieve these requirements, which are outlined in the activity diagram. Lastly, the interactions of the tool components and each of the previous diagrams are depicted in the deployment diagram.

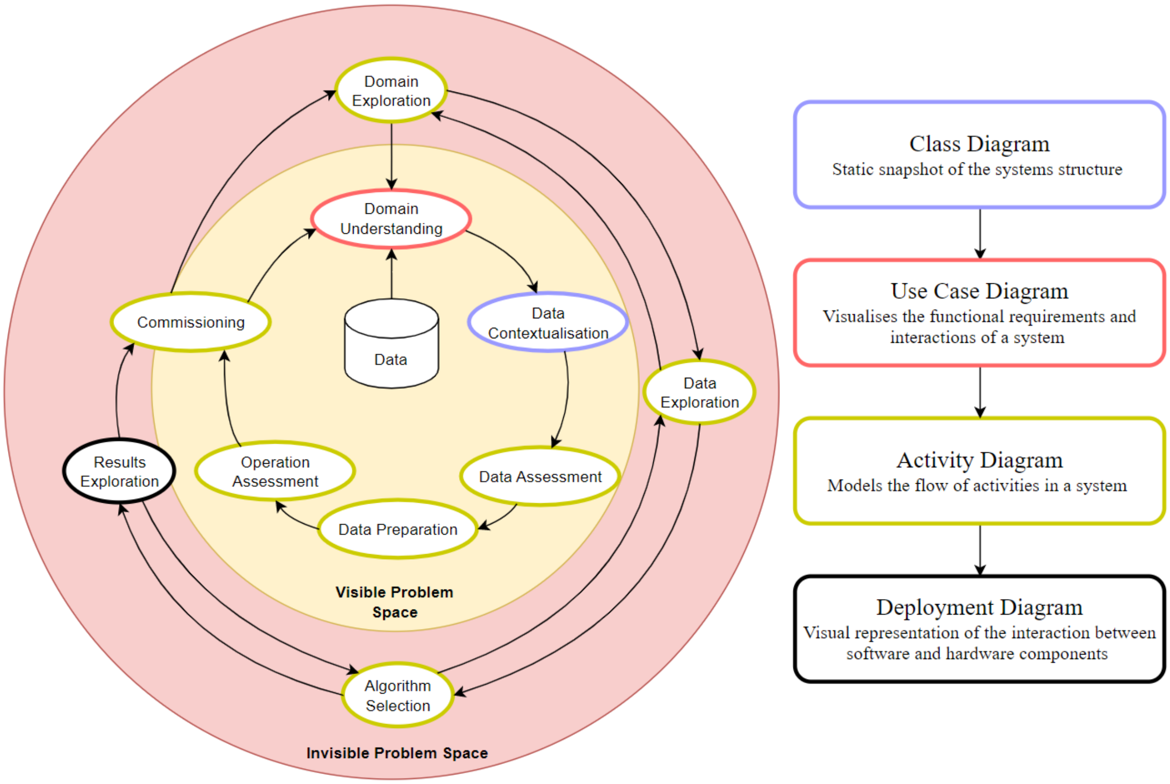


Figure 6-2 Overview of the IDAIC phases associated with the UML diagrams

6.2.1 Class Diagram

Firstly, a class diagram is developed to show the structure of the software tool in terms of classes, objects, and their relationships. In this diagram, the classes represent the templates or blueprints that define the objects. Meanwhile, the objects are the specific instances created from the classes that represent individual entities within the system. Taking a bakery as an analogy, the class would be the recipe to bake a cake, the object would be the particular type of cake such as chocolate, while the instance would be the individual cake that is baked using the recipe, or class. This is similar to the resource description framework (RDF) data models that are used to convey the metadata information in Project Haystack [203] and Brick [204] ontological relationships. Project Haystack refers to RDF being considered as the “lingua franca” in this space – a language systematically used to make communication possible between people not sharing a first language. Therefore, the class diagram is most appropriately used to describe the data model in this context rather than the structure of the tool. An overview of the data model is shown in Figure 6-3 for a subset of three air temperature sensors. Given the development of this tool is part of the DENiM project with multiple pilot sites involved, the model begins with identifying the case study site. Then within this site, several different systems may be defined. While the DENiM project has multiple systems within its scope, the HVAC system is the area of focus for this research. Within the system, multiple assets or equipment may be defined such as the AHU. Finally, the data points relating to the AHU are defined. The data mapping file utilised in the data contextualisation phase of the framework implementation in Chapter 5 is extended to capture this data architecture. Furthermore, the data is contextualised by defining the metadata, such that the data points in Figure 6-3 are air temperature sensors. Therefore in this context, the class diagram has aided the deployment of the data contextualisation phase as shown in Figure 6-2.

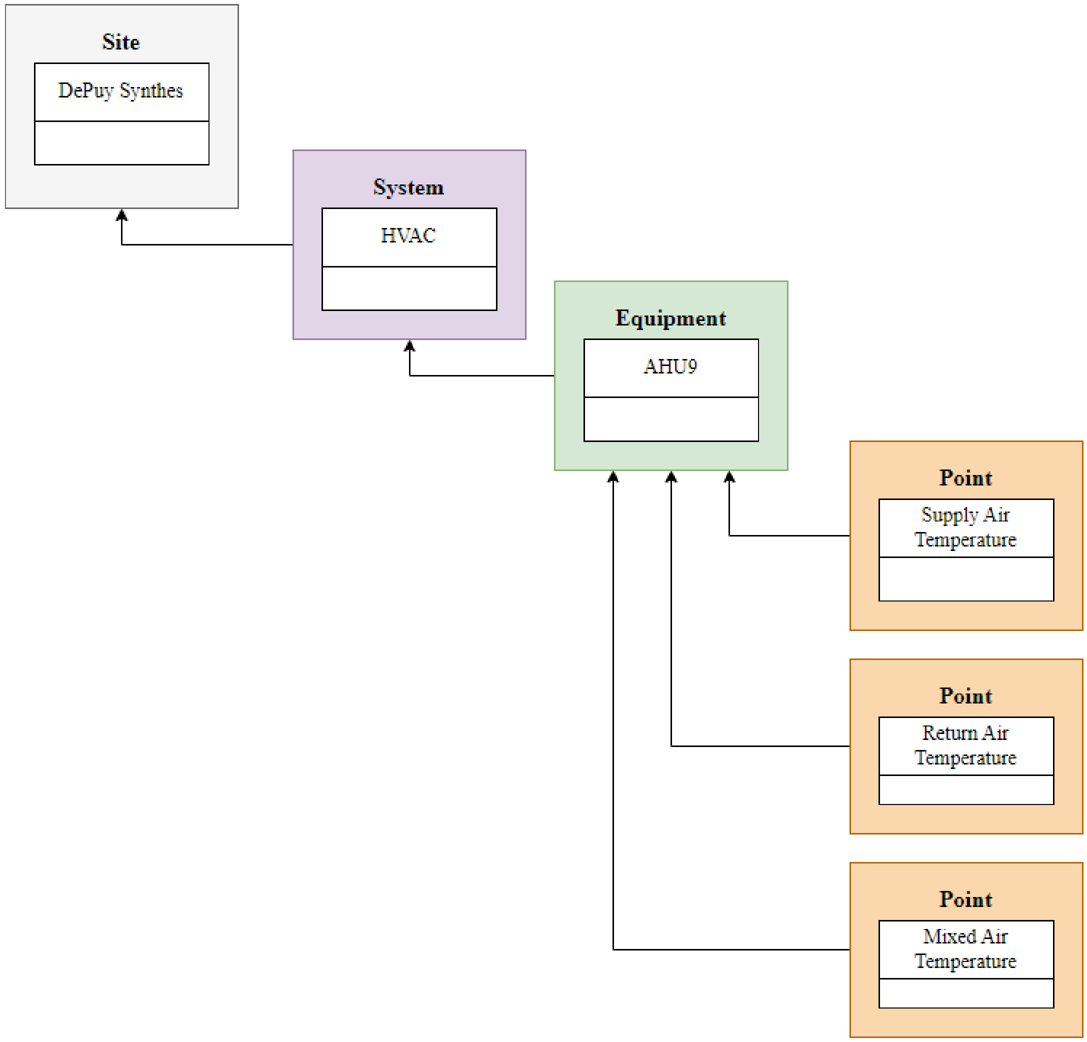


Figure 6-3 Class Diagram of the data model

6.2.2 Use Case Diagram

Secondly, the needs of the users of the tool must be considered. This is a key consideration in the Industry 5.0 paradigm which is associated with the domain understanding phase of the IDAIC framework. Three key users are considered; an energy manager, a facilities technician and a commissioning engineer. These users are selected to embody the three levels of data-driven FDD as outlined in the framework implementation. The requirements of each user were elicited through the weekly stakeholder meetings, where representatives of each user group participated. Open discussions were held in these meetings where participants discussed their current workflows and challenges, along with their expectations of the tool. Additionally, follow-up interviews were conducted to ensure the elicited requirements satisfied the user's needs. At the first level, daily energy consumption analysis, the energy manager will be driven by reducing the overall consumption of the AHU. At the second level, precision analysis and stability analysis, the facilities technician will be driven by ensuring the AHU is achieving the designed air temperature set points. While at the third and final layer, data-driven modelling, the commissioning engineer will be driven by detecting the faulty sensors and components in need of replacement. In UML, these users are named actors which are defined as a user that participates in a use case to accomplish a purpose [199]. Each of these actors will have different requirements and therefore the tool must fulfil multiple needs to be considered effective. For example, a facilities technician or commissioning engineer will require operational data to be displayed. While an energy manager will require high-level quantification of the opportunity for energy consumption reduction to decide when the maintenance activity should take place. The specific requirements of these three actors were elicited through the DENiM stakeholder meetings and are outlined in Table 6-1.

Table 6-1 Table of actor requirements

	Requirements	Energy Manager	Facilities Technician	Commissioning Engineer
1	Identify/Display faults as well as the reason for the fault.	X	X	X
2	Display AHU performance by best/worst in terms of the number and cost of faults.	X		
3	Identify/Display the actual versus ideal operation of an individual AHU.	X		X
4	Identify/Display the top ten issues across each AHU by cost or criticality.	X		
5	Display equipment operating data		X	X

Table 6-2 Table of software requirements

	Requirements
1	Access the data from the BMS
2	Contextualise the raw data
3	Store the contextualised data in a standardised manner to enable the data to be easily queried for visualisation
4	Assess the data for broken, bad-quality and background issues
5	Apply logic to uncover visible faults
6	Apply logic to uncover invisible faults

Furthermore, the requirements of the software must also be considered. For example, the tool must be able to access the data and it requires the FDD logic to uncover the faults. The requirements of the software are outlined in Table 6-2. Both the requirements of the users and the requirements of the software must be clearly outlined to ensure they are fulfilled. This may be achieved by translating the requirements into system functionality through use cases. The use cases provide a brief overview of the requirement and the steps that must be completed to fulfil it. The requirements outlined in Table 6-1 and Table 6-2 are therefore translated to a use case in Table 6-3.

Table 6-3 Use Case Overview

		User Requirements in Table 6-1					*Software Requirements in Table 6-2*					
	Use Case	1	2	3	4	5	1	2	3	4	5	6
1	Data Collection (Provided in Table 6-4)						X					
2	Data Mapping							X				
3	Data Pre-processing								X			
4	Data Assessment									X		
5	Identify Visible Faults										X	
6	Explore Invisible Faults											X
7	Check Operation versus Ideal Control Strategy			X								
8	Display Fault Information & Diagnosis	X	X									
9	Display Top Ten Faults/Costs		X		X							
10	Display Equipment Operating Data					X						

The interactions between the actors and the tool in terms of the specified use cases are then diagrammatically depicted in Figure 6-4. This graphical representation ensures that each user can interact with the tool to satisfy their specific needs. This highly interpretable and documented means of conveying the user's requirements proved to be a key diagram. During the tool development process, internal restructuring on the industrial partner site occurred which caused a change in personnel for the user roles. This diagram aided the new personnel to quickly understand the aims of the tool given the documented input from the previous user.

Each of the use cases defined in Table 6-3 must then be translated into a series of steps that must be completed to achieve the requirement. For example, the steps that must be achieved by the tool for the data collection use case are outlined in Table 6-4. In this use case, the tool must establish a connection with the industrial site. Then the tool determines the required data points from the BMS mapping, which has its own use case as in Table 6-3. The desired AHU and sensors are then used to identify the required data which is then read by the tool. In the case that the tool cannot collect the data, an exception is raised which logs an error.

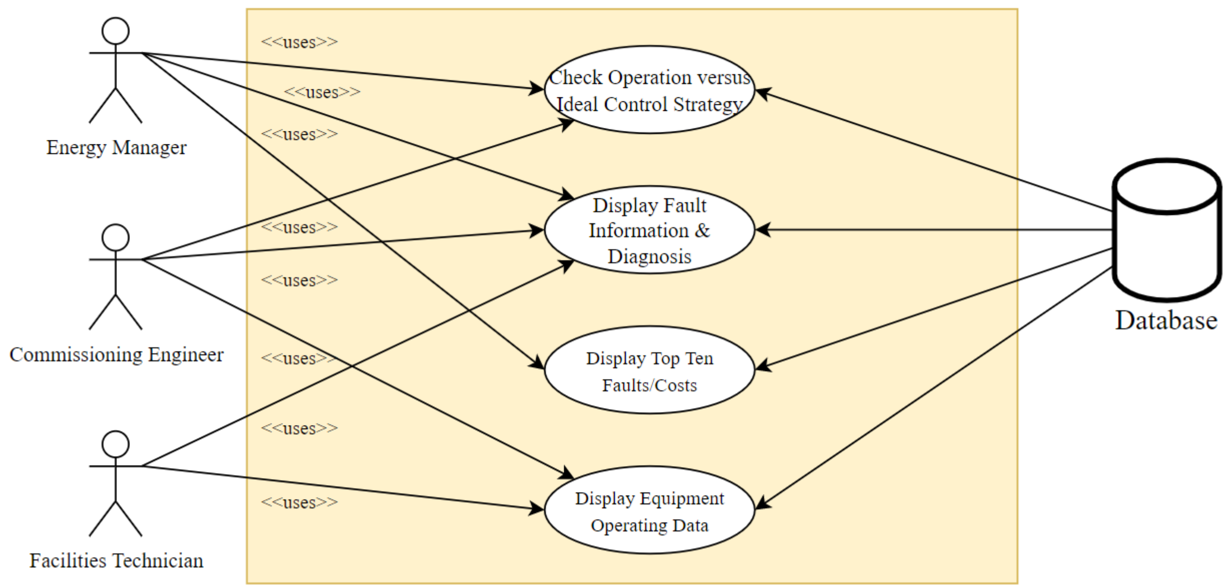


Figure 6-4 Use Case Diagram

Table 6-4 Sample use case for data collection

Name Use Case 1	**Data Collection**
Summary	Collect data daily from the BMS for the AHU
Primary Actor	Software
Requirement	Software Requirement 1
Results	Data accessed
Normal Flow	
Step	Action
1	Get the site connection information
2	Identify BMS Mapping including the number of redundancy required
3	Identify desired AHU
4	Identify Unit Sensors
5	Identify required data
6	Read data
	Exceptional Flow
Step	Action
6	Unable to open data store so log message

6.2.3 Activity Diagram

Thirdly, the activity diagram in Figure 6-5 provides a flowchart of the users' behaviour with the tool through the data assessment, data preparation, operation assessment, commissioning, domain exploration, data exploration and algorithm selection phases of the IDAIC framework. Given that the implementation of the framework in Chapter 5 did not extend to the algorithm selection phase, further guidance is now provided for choosing the appropriate approach in the sequence of activities that are outlined as follows.

The feature level and multi-feature level data assessment logic are used to detect sensor issues, while the steady-state detector is used to uncover transient control issues. The outputs of the data assessment component are then used to inform the preparation of the data for the application of the knowledge-based FDD. For example, this may involve applying smoothening techniques to the data during transient state conditions. There are then three possible outcomes of the operation assessment, visible fault-free, undiagnosed faulty operation and diagnosed faulty operation. Firstly, for visible fault-free operation, a baseline may be established which enables the detection of the subtle degradation-type invisible faults. Based on the selection process proposed by Matetic et al [105], the exploratory unsupervised analysis may take the form of finding groups or rules which lead to clustering and association rule mining, respectively [105]. Secondly, the undiagnosed faulty operation requires further investigation of the root cause of the fault. In well-established systems, previous faults that have occurred may be compared to the incoming data to align the undiagnosed fault with a past fault in a supervised classification approach. Meanwhile, in less established systems, the impact or severity of the fault may be estimated in a regression approach to quantify the impact of the fault on the AHUs operation. This would quantitatively justify conducting maintenance checks on multiple potential causes when a specific diagnosis is not available.

This activity, outlined in Figure 6-5, imitates the behaviour of the three identified actors. Firstly, the regression analysis for the impact of a visible fault prompts an energy manager to rectify increased energy consumption. Secondly, the facilities technician is provided with a fault classification based on previous faults. For example, a fault may be classified as heating coil fouling which would then identify a heating coil in need of cleaning. Thirdly, a commissioning engineer would be provided with detailed analysis and evidence for the occurrence of invisible faults during commissioning. This would reduce the likelihood of missed opportunities for improvement during commissioning.

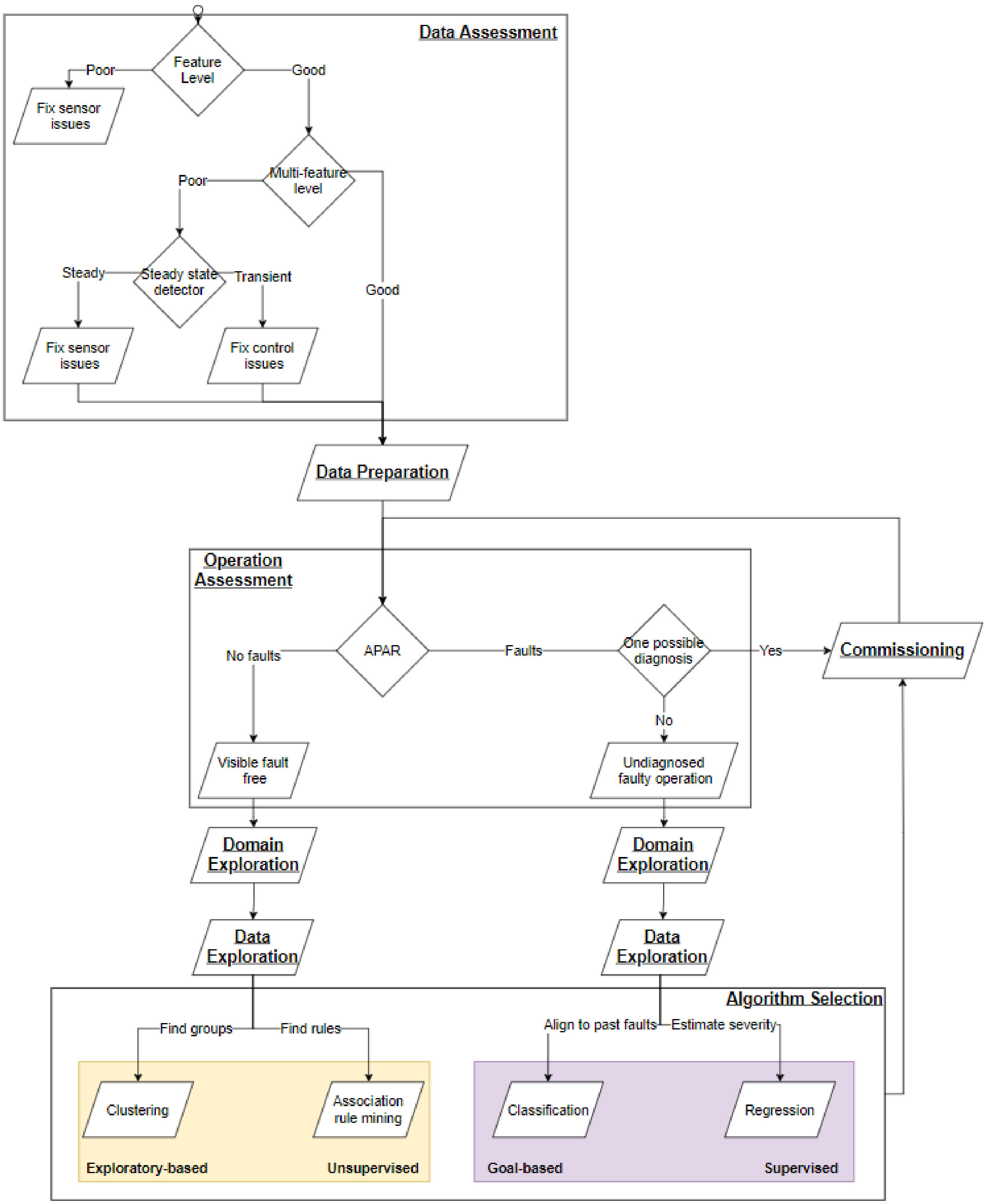


Figure 6-5 Activity Diagram

6.2.4 Deployment Diagram

Lastly, the deployment diagram outlines how the tool is deployed and executed in the real world. Therefore, the final phase of the IDAIC framework, results exploration, may be linked to the physical development of the graphical user interface (GUI) that enables the actors to fulfil their requirements. As this final diagram relates to the remit of other partners in the DENIM EU project in terms of the selection of software tools to fulfil the use cases, it is outside the scope of this research. However, as this work contains critical aspects to the development of the tool it is worthy of inclusion. Therefore, the main components are outlined in the deployment diagram in Figure 6-6. This diagram shows both the key steps and software elements in the tool development process which is then split into three regions or layers; the data layer, the processing layer and the graphical user interface layer.

The outcome of this tool design subsection is a set of related UML diagrams that aim to facilitate discussion between tool developers and tool users. An overview of the relationship between the IDAIC phases, UML diagrams and uses cases (requirements of the tool) is outlined in Table 6-5. Given the non-process nature of UML, the IDAIC framework offers guidance for establishing connections between the UML diagrams. This guidance has led to the development of a tool development pipeline, as depicted in Figure 6-6. Therefore, the IDAIC framework has provided the guidance to sequentially order the use cases to be fulfilled through the tool pipeline. This is further discussed in the following subsection.

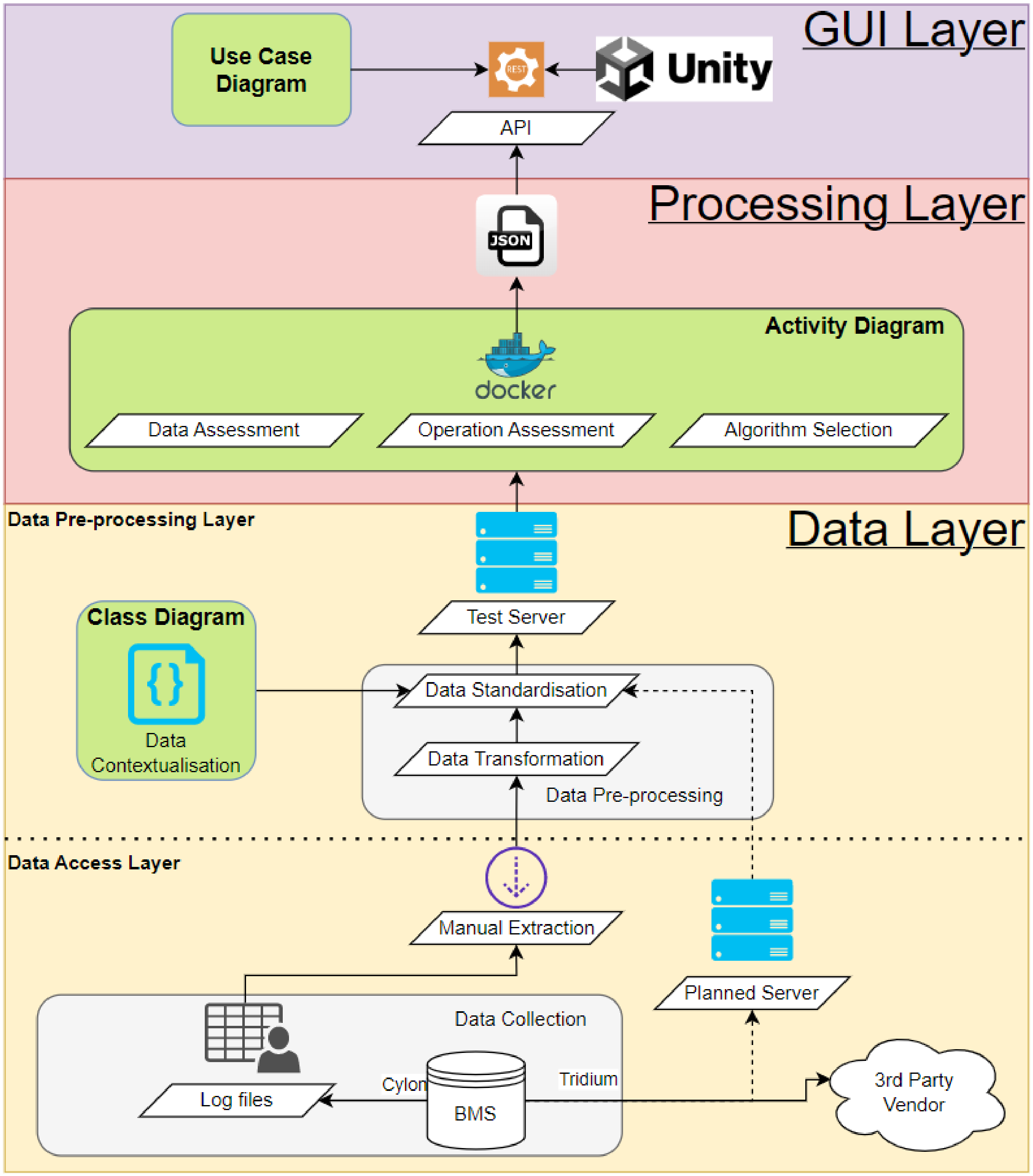


Figure 6-6 Deployment Diagram

Table 6-5 Overview of the relationship between the IDAIC phase, UML diagram, and use cases

IDAIC Phase	**UML Diagram**	**Use Cases**
Domain Understanding	Use Case Diagram	Outlines software requirements (1-6) and user requirements (7-10)
		Fulfils Data Collection (1)
Data Contextualisation	Class Diagram	Fulfils Data Mapping (2)
Data Assessment	Activity Diagram	Fulfils Data Assessment (4)
Data Preparation		Fulfils Data Pre-processing (3)
Operation Assessment		Fulfils Identify Visible Faults (5)
Commissioning		
Domain Exploration		Fulfils Explore Invisible Faults (6)
Data Exploration		Fulfils Display Equipment Operating Data (10)
Algorithm Selection		Fulfils Check Operation vs Ideal Control Strategy (7)
Results Exploration	Deployment Diagram	Fulfils Display Fault Information & Diagnosis (8) Fulfils Display Top Ten Faults/Costs (9)

6.3 Tool Implementation

A simplification of the process of implementing or deploying the tool outlined in Figure 6-6 is shown in Figure 6-7. The first layer, the data layer, involves both accessing and pre-processing the raw data. The second layer, the processing layer, involves the application of methods to fulfil the requirements that are visualised in the final layer, the GUI layer. The contributions of the DENiM research group are outlined in relation to the software engineering-based activities in the back-end (data layer and processing layer) and the front-end (GUI layer). The contribution of this research is also outlined, which concerns the logic behind the development of the tool. This process is discussed in further detail in the following subsections.

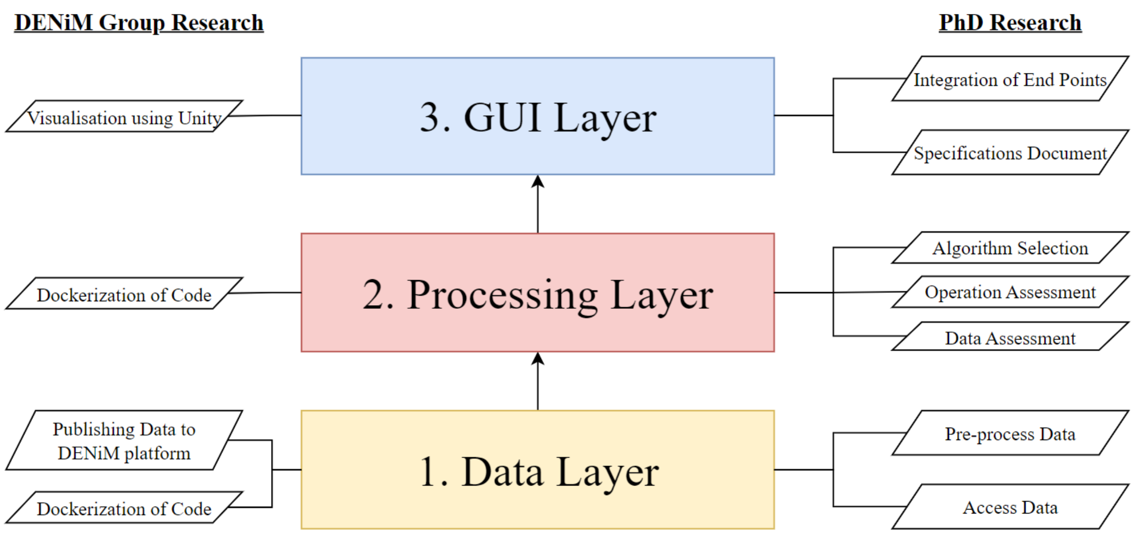


Figure 6-7 Tool implementation process and research delineation

6.3.1 Data Layer

The data layer is concerned with accessing the raw data and storing it in an easily accessible way. It is split into two parts; data access and data pre-processing. In the data access sublayer, the domain understanding phase is needed to guide communication between software engineers and domain experts so that data may be appropriately collected. As discussed in the data contextualisation phase of the framework implementation, there are two main avenues for accessing the AHU data. The Cylon Unitron log files only update every 24 hours while the Tridium JACE box updates every

15-min. Therefore, collecting the data via the Tridium system is preferred for near-real-time capabilities. However, the connection made by the third-party vendor cannot be replicated due to the site's data security policy. Therefore, an internal server is currently under development to extract the data from the BMS to the site's secure and easily accessible data historian. While this server project is ongoing, the Cylon Unitron log files may be used as they essentially represent the same data that is available through Tridium. Therefore, storing this data in a "sandbox server" - a virtual environment to build, test and deploy software [205] - would offer a short-term solution to mirror the site's future internal server. This would enable a modular approach to the data collection use case so that the development of the tool may be continued, as shown in Figure 6-6. This is an example of an unforeseen practical consideration that the modular nature of the IDAIC framework was able to address.

Given that the data may now be accessed, it must be appropriately prepared for analysis in the data pre-processing layer. In this sublayer, the activities related to the pre-processing of data are performed, namely; data mapping and data pre-processing. The data mapping process outlined in the framework implementation is extended to align with the class diagram of the data model. This involves extending the naming convention to a hierarchical structure that describes the relationships between the data. This allows the tool to easily query the data points that are related to the AHU that is selected by a user. Furthermore, the metadata is recorded so that the tool may correctly interpret the data as an air temperature sensor that is measured in degrees Celsius for example. This satisfies one component of the data pre-processing step, the standardisation of the data.

The other component of the data pre-processing step is data transformation. Given that the log files store data in a bespoke manner as in Figure 5-6, this is a critical step to ensure that the correct value is assigned to the correct timestamp. While this step is only required for the Cylon Unitron log files, it is another practical consideration that may not be easily overcome by a pre-built commercial offering. The structured approach to pre-process the log files was conveyed to the DENiM partners that are responsible for the software used to develop the tool. The versatile and general-purpose nature of the Python code enabled the complex and bespoke data storage process to be conveyed to software engineers with no prior knowledge of AHUs or this form of data collection.

6.3.2 Processing Layer

Once the data has been appropriately prepared and stored in the test server, it is ready to be analysed to detect data issues, visible faults and invisible faults. This analysis is conducted on the DENiM Platform which is maintained by the software engineering partners on the DENiM EU project. Therefore, strong collaboration was needed to communicate the data assessment, operation assessment and algorithm selection logic. To facilitate collaboration with research partners, weekly meetings were scheduled to discuss the translation of code developed in the implementation described in Chapter 5 into a Docker file. A Docker file is a text file that contains the commands to be carried out, which are then containerised so that they may be executed on any infrastructure [206].

As in Figure 6-6, there are three main components; data assessment, operation assessment and algorithm selection. Firstly, the data assessment logic is implemented using the novel data assessment decision tree. This provides the commissioning engineer with an overall health status of the data before investigating any faults that may have been triggered. This involved the creation of new data points to output the results of the data assessment. The data point is contextualised in a manner similar to that of the sensor data. For example, the mixed air temperature sensor (MaTemp), includes information about the site location (P1 for pilot site 1), the equipment (AHU which captures that the sensor belongs to an air handling unit which is part of the HVAC system) and the identifier (09 which indicates AHU9). The data generated at the single feature level is aligned to this sensor by extending the name to include “availability” for missing data, “spike” for detecting outliers and “stuck” for detecting values that do not change. Meanwhile, the multi-feature level conducts a comparison check of multiple features which is not linked to the data points but to the AHU. For example, the P1, AHU, and 09 are linked to the semantic consistency rule. This then creates a data point that outputs the detection of drift, bias and precision degradation faults for the defined similar sensors. Similarly to the semantic consistency rules, a steady-state detector is defined for the AHU.

Fault data named in accordance with class diagram

```
# Mode 2: Cooling with Outdoor Air
df['P1_AHU9_rule_05'] = 0
df.loc[(df['Mode'] == "2") & (
        (df["fields.p1-ahu9-oatemp-01"] > 21 - DeltaSF + e_t) | (df["fields.p1-ahu9-oatemp-01"] > 20 - DeltaSF + e_t)), 'P1_AHU9_rule_05'] = 1
df['P1_AHU9_rule_05_c'] = 0
df.loc[(df["Mode"] == "2_c") & ((df["fields.p1-ahu9-oatemp-01"] > 21 - DeltaSF + e_t) | (
        df["fields.p1-ahu9-oatemp-01"] > 20 - DeltaSF + e_t)), 'P1_AHU9_rule_05_c'] = 1

df['P1_AHU9_rule_06'] = 0
df.loc[(df['Mode'] == 2) & (df["fields.p1-ahu9-datemp-01"] > df["fields.p1-ahu9-ratemp-01"] - DeltaRF + e_t), 'P1_AHU9_rule_06'] = 1
df['P1_AHU9_rule_06_c'] = 0
df.loc[(df["Mode"] == "2_c") & (df["fields.p1-ahu9-datemp-01"] > df["fields.p1-ahu9-ratemp-01"] - DeltaRF + e_t), 'P1_AHU9_rule_06_c'] = 1

df['P1_AHU9_rule_07'] = 0
df.loc[(df['Mode'] == "2") & (abs(df['fields.p1-ahu9-datemp-01'] - DeltaSF - df["fields.p1-ahu9-matemp-01"]) > e_t), 'P1_AHU9_rule_07'] = 1
df['P1_AHU9_rule_07_c'] = 0
df.loc[
    (df["Mode"] == "2_c") & (abs(df['fields.p1-ahu9-datemp-01'] - DeltaSF - df["fields.p1-ahu9-matemp-01"]) > e_t), 'P1_AHU9_rule_07_c'] = 1
```

Figure 6-8 Extract from the Docker file for the operation assessment (APAR Mode 2)

Secondly, for operation assessment, the knowledge-based FDD logic is applied using the extended APAR rules developed in Chapter 5 to identify the visible faults. An extract from the Docker file is shown in Figure 6-8, which shows how the data is contextually named in accordance with the class diagram. The explanation of the triggered rule along with the possible diagnoses is provided in Figure 6-9 and Figure 6-10. This contextualises the rules for users that are not familiar with AHU FDD. The rows highlighted in red are purposely not included because; the changeover temperature is not fully applicable given zones with differing set points, the data granularity is not sufficient to analyse mode transitions per hour and rule 29 is undefined so that temperature difference across coil faults are conveniently numbered 30 or above. These are additional rules that have been added to check for leaking valves. Further insight and energy consumption quantification are then provided by the third component. The data-driven algorithm is applied by creating an energy baseline using the method described in subsection 5.3.2. Using domain knowledge to create a proxy indicator of the temperature difference across the heating coil for heating energy consumed, a simple linear regression model is fitted to the data. In Figure 6-11 the daily heating energy consumed by the AHU is plotted against the heating degree days obtained from a nearby weather station. Heating degree days is a measure of how much, and how long, the outside air temperature was lower than a specific base temperature (15.5°C) [207]. The difference between the heating energy consumed during COVID operation (red) and normal or baseline operation (blue) is shown in Figure 6-11. As discussed in Chapter 5, the failure to return this AHU to normal operation in line with the other AHUs on site is an example of an invisible fault. The data-driven regression-based FDD using energy consumption is an effective means of detecting the occurrence of invisible faults that have an impact on the energy consumed by the AHU.

To validate that the logic was correctly applied before putting it in the Docker file, the code was tested on the sandbox server. This shared cloud-based platform enabled the validation of the code daily, using the previous 24 hours of data. This allowed a collaborative approach to be taken to the tool deployment, ensuring that the intricacies of the data assessment, operation assessment and algorithm selection were correctly applied.

Fault	Impact				Possible root cause																					Fault Explanation
Rule No.	Energy	Maintenance	Indoor Air Quality	Comfort	Supply Air Temperature Sensor Error	Return Air Temperature Sensor Error	Mixed Air Temperature Sensor Error	Outdoor Air Temperature Sensor Error	Leaking cooling coil valve	Stuck cooling coil valve	Undersized cooling coil	Fouled cooling coil	Chilled water supply temperture too high	Chilled water circulating pump fault	Chilled water not available	Leaking heating coil valve	Stuck heating coil valve	Undersized heating coil	Fouled heating coil	Hot water supply temperature is too low	Hot water circulating pump fault	Leaking mixing box damper(s)	Stuck mixing box damper(s)	Controller tuning error	Controller logic error	
1	x			x	x		x		x	x							x	x		x	x					*Supply air temperature is less than mixed air temperature in heating mode*
2	x		x			x	x	x														x	x			*Fraction of outdoor air entering the AHU is either too high or too low*
3	x			x	x				x	x							x	x	x	x	x			x		*Insufficient heating. System is out of control*
4	x			x	x				x	x							x	x	x	x	x			x		*Insufficient heating. System is out of control*
5	x				x			x																		*Outdoor air temperature is too high for this mode*
6				x	x	x																				*Inconsistency between supply and return air temperatures*
7	x				x		x		x	x						x	x									*Inconsistency between supply and mixed air temperatures*
8	x				x			x								x	x					x	x			*Outdoor air temperature is too low for mechanical cooling with 100% outdoor air*
9	x																								x	*Outdoor air temperature is too high for this mode*
10	x						x	x														x	x			*Inconsistency between the outdoor and mixed air temperatures*
11	x			x	x		x			x	x		x	x		x	x									*Inconsistency between the supply and mixed air temperatures*
12				x	x	x				x	x	x	x	x	x	x	x							x		*Inconsistency between the supply and return air temperatures*
13	x			x	x					x	x	x	x	x	x	x	x							x		*Insufficient cooling. System is out of control*
14	x			x	x					x	x	x	x	x	x	x	x							x		*Insufficient cooling. System is out of control*
15	x																								x	*Outdoor air temperature is too low for this mode*

Figure 6-9 Knowledge-based FDD rules explanation and possible diagnosis

Fault	Impact				Possible root cause																					Fault Explanation
Rule No.	Energy	Maintenance	Indoor Air Quality	Comfort	Supply Air Temperature Sensor Error	Return Air Temperature Sensor Error	Mixed Air Temperature Sensor Error	Outdoor Air Temperature Sensor Error	Leaking cooling coil valve	Stuck cooling coil valve	Undersized cooling coil	Fouled cooling coil	Chilled water supply temperture too high	Chilled water circulating pump fault	Chilled water not available	Leaking heating coil valve	Stuck heating coil valve	Undersized heating coil	Fouled heating coil	Hot water supply temperature is too low	Hot water circulating pump fault	Leaking mixing box damper(s)	Stuck mixing box damper(s)	Controller tuning error	Controller logic error	
16	x			x	x		x			x	x		x	x		x	x									*Inconsistency between the supply and mixed air temperatures*
17				x	x	x				x	x	x	x	x	x	x	x							x		*Inconsistency between the supply and return air temperatures*
18	x		x			x	x	x														x	x			*Fraction of outdoor air entering the AHU is either too high or too low*
19	x			x	x					x	x	x	x	x	x	x	x							x		*Insufficient cooling. System is out of control*
20	x			x	x					x	x	x	x	x	x	x	x							x		*Insufficient cooling. System is out of control*
21	x																								x	*Cooling coil valve, heating coil valve, and mixing box dampers are modulating simultaneously*
22	x																								x	*Cooling coil valve and heating coil valve are modulating simultaneously*
23	x																								x	*Heating coil valve and mixing box dampers are modulating simultaneously*
24	x																								x	*Cooling coil valve and mixing box dampers are modulating simultaneously*
25				x																				x		*Supply air temperature set point is not being satisfied*
26	x					x	x	x																		*Mixed air temperature is less than the minimum of return and outdoor air temperatures*
27	x					x	x	x																		*Mixed air temperature is greater than the maximum of return and outdoor air temperatures*
28	x	x																						x		*Excessive mode transitions occuring*
29																										*Purposely left blank*
30	x	x																								*Temperature rise across the heating coil when the valve has been closed for 1 hour*
31	x	x																								*Temperature drop across the cooling coil when the valve has been closed for 1 hour*
32	x	x																								*Temperature rise across the plantroom reheat coil when the valve has been closed for 1 hour*
33	x	x																								*Temperature rise across the waxroom reheat coil when the valve has been closed for 1 hour*

Figure 6-10 Knowledge-based FDD rules explanation and possible diagnosis continued

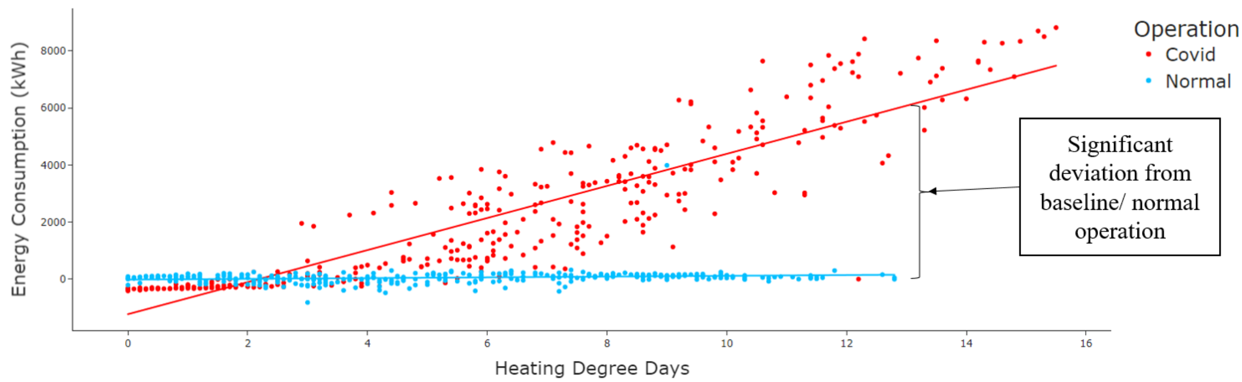


Figure 6-11 Data-driven FDD visualising difference between regression line for COVID_Operation vs baseline normal operation

6.3.3 Graphical User Interface Layer

The requirements and functionality of the tool are then visually delivered in the graphical user interface (GUI) layer. The development of the GUI was led by the coordinating institute of DENiM, MTU, with the requirements and desired functionality of the tool provided by this research. The tool design using UML, outlined in subsection 6.2, was documented in the form of requirements specifications. This requirements specification outlined the data structure using the class diagram, the requirements of the actors using the use case diagram, the desired layout of GUI using the activity diagram, and the tool's structure using the deployment diagram. Additionally, graphs and plots from the framework implementation were provided as a static prototype of the desired GUI. An extract from the requirements specification is outlined in Table 6-6. The first level of the data-driven FDD approach is to analyse the overall consumption of the AHU and identify anomalies. Table 6-6 provides an overview of logic behind the development of the energy performance indicator (EnPI) in Figure 6-11, named the energy consumption anomaly (ECA), for the GUI developers to replicate in the software environment.

Table 6-6 Definition of the energy performance indicator (EnPI)

Name	Energy Consumption Anomaly
Description	Using the available sensor data, the electrical and hydronic energy consumed is compared to the degree day baseline.
Scope	The EnPI is relevant to the energy consumed by the AHU
Formula	**Energy Consumption Anomaly (ECA) =** $(\text{Actual kWh} - \text{Expected kWh})$ Actual kWh = Electrical Energy + Hydronic Energy Expected kWh = Obtained from kWh consumed vs degree days Electrical Energy (Return and Supply air fan): • Direct meter reading, or; • Cube law (requires fan speed) Hydronic Energy (Hot and Chilled water): • $Q = mc_p\Delta T$* • For consistency, Q is converted to electrical units by dividing by an assumed chiller COP for cooling, and multiplying by an assumed boiler efficiency for heating. ***This is an air-side calculation. Where m is the supply air mass flow rate, c_p is the specific heat capacity of the supply air and ΔT is the air temperature difference across the relevant heating or cooling coil**
Unit of measure	kWh
Range	Dependent on the size of the AHU
Trend	Closer to zero is best (normal operation) The higher, the more energy consumed/ inefficient The lower, the less energy consumed (perhaps wrongly in relation to thermal comfort)
Timing	Depending on the nature of the data acquisition, results may be: • Real-time - after each new data acquisition event • On demand - after a specific data selection request • Periodically - done at a certain interval, e.g. once per day
Audience	Audience is the user group typically using this EnPI. The user groups used in this part of • Facilities Technicians – personnel responsible for the direct operation of the equipment • Commissioning Engineers – personnel responsible for carrying out the maintenance activities • Energy Manager – personnel responsible for the overall energy consumption
Notes	Assumptions: • Access to degree days data from a local weather station • c_p of supply air is 1.006 kJ/(kgK) • m of supply air is density (1.225kg/m) multiplied by volumetric flow rate

In consultation with the key stakeholders that were informed by the requirements specification document, three main screens, or levels, were developed; screen 1 site overview, screen 2 AHU overview and screen 3 component analysis. At the site level, the requirements of the energy manager are satisfied. The ECA for the AHU is provided along with the most costly faults which aids the energy manager to prioritise maintenance activity. At the AHU level, an overview of the specific AHU is provided for a facilities technician as in Figure 6-12. This level gives an overview of the number of visible faults occurring in a particular AHU, along with the possible diagnoses and trend data analysis. The component level provides greater detail that would be required by a commissioning engineer, such as data assessment metrics and comparison of the current performance of the AHU and its components to their baselines.

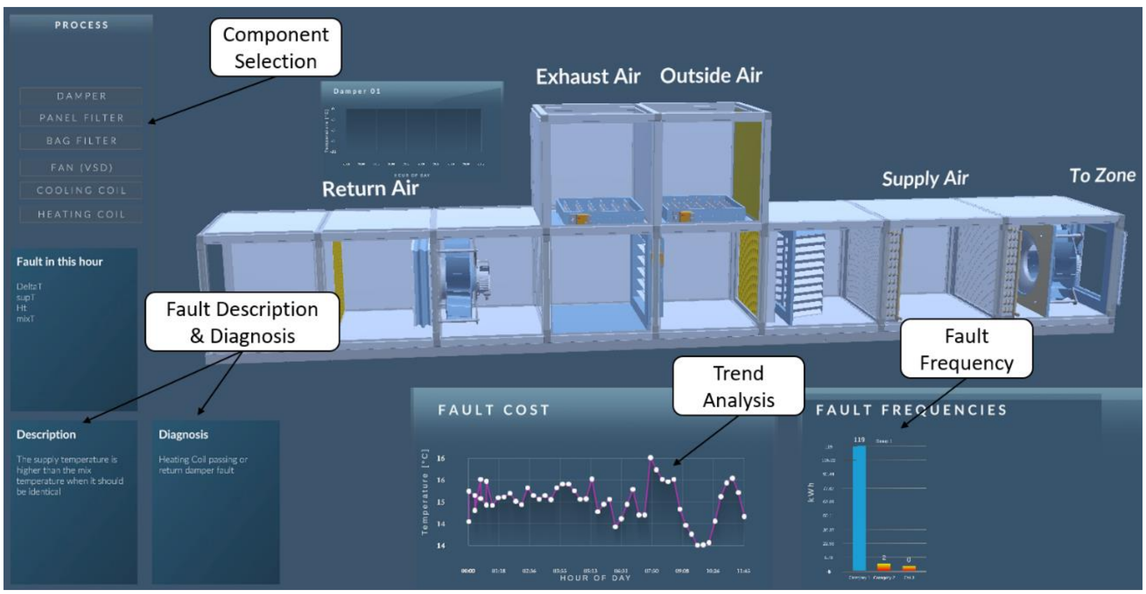


Figure 6-12 Example of the AHU level of the GUI

6.4 Results and Evaluation

The implementation of the tool in the industrial site uncovered four key practical considerations that needed to be addressed; data access, resource constraints, cyber security concerns and latency issues.

Firstly, accessing the data was a significant issue considering the BMS is hosted on a physical PC running Windows XP Professional. This is an example of legacy equipment that the industrial site is attempting to update to reap the benefits of digitalisation. However, this is a long-term project that posed a risk to the tool development should it not be completed on time. As mitigation, a parallel track was created to access the data as shown in the data layer in Figure 6-6. The log files that were used in the IDAIC implementation in Chapter 5 were used to simulate the data that would be available after the BMS upgrade project was completed. Given that IDAIC is a framework-type solution it allowed the flexibility to deviate from the intended data acquisition method of the partners on the DENiM EU project. This is an example of how the IDAIC framework can address practical issues that contribute to the practice-theory gap.

Secondly, resource limitations are another key consideration. In this context, the resources refer to both the availability of the industrial site employees and finances for tool deployment. During the course of the tool deployment, there was high turnover on the industrial site which required experienced personnel to be assigned to critical day-to-day operations. This resulted in a reduced availability of site-specific domain expertise to provide context to the EU project team. This issue further highlights the unsustainable nature of solely relying on domain expertise as discussed in 2.3.1. However, while the availability of human resources was reduced, the tool development continued as the site-specific knowledge was effectively elicited through the UML approach that was underpinned by the structured IDAIC framework. This ensured that the requirements of the users were fulfilled which also contributed towards addressing the financial limitations. As the tool highlights the most common faults and the assumed cost of their occurrence, greater justification is provided to overcome the hesitancy of investing in maintenance activity.

Thirdly, the cyber security threat is a significant concern in industry, especially for multinational organisations. This is linked to data privacy concerns and the threat that malware (malicious software) has to the viability of an organisation. This was a very pertinent issue during the tool deployment as the DENiM coordinators, MTU, suffered a cyber-attack in February 2023 [208]. This caused significant delays to the tool's development as the MTU systems were non-operational while the attack was investigated. Therefore, the preferred approach of this site is to host any software on its internal platforms which are protected behind the site's firewalls. Therefore, the software agnostic approach of the IDAIC framework provides a practical in-house solution for industrial sites that are uncomfortable allowing third-party tool providers live connection to their systems.

Lastly, the front-end GUI developer on the EU project experienced some latency issues when fulfilling the user requirements on the Unity platform. This was caused by the high demands of both rendering a 3D model of the AHU and visually graphing many of the data points. This issue was overcome by splitting the GUI layer so that the Unity platform rendered the 3D model of the AHU and high-level visualisation, while another visualising tool was used for more detailed analysis needed for uncovering the invisible faults. This is a further example of how the structured, yet flexible, nature of the IDAIC approach enables practical amendments where necessary.

6.5 Conclusion & Future Work

The purpose of this chapter was to determine if the IDAIC framework was capable of dealing with the practical considerations that arise through tool deployment in a case study. The development of the tool in the presence of data access issues, resource limitations, hesitancy regarding cyber security threats and latency issues is proof of IDAIC's effectiveness. The tool enables the continuous monitoring of the AHU needed for proactive maintenance which contributes to improved efficiency and reaching climate targets.

The DENiM EU project is ongoing with an end date of October 31st 2024. Therefore, further development of the tool will occur which will include the addition of other AHUs, refinement of the GUI, and the addition of other HVAC components such as chillers.

7 Chapter 7 – Conclusions & Future Work

7.1 Summary of Research

This research aimed to contribute to the Industry 5.0 paradigm shift through the development and implementation of a framework that aids the transition to proactive maintenance approaches on air handling units in the industrial setting. The review of the research relevant to this area elicited four distinct research gaps. Firstly, the knowledge gap described the difference between the reactive engineering mind-set and the proactive data science mind-set. Secondly, the data gap described the difference between high-quality experimental data in the literature and flawed real-world data in industry. Thirdly, the operational gap described the difference between the known operating states of AHUs in experimental studies whereby faults are abruptly enforced, and the unknown operating states of AHUs in the industrial setting whereby unnoticed degradation occurs. Lastly, the practice-theory gap described the difference between the predominantly data-driven approach in literature and the predominantly rule-based approach in commercial settings. This then leads to the main research question, *"How can reproducible and usable data-driven approaches be developed, implemented, and deployed to reduce energy consumption in the industrial setting through the transition to proactive maintenance of AHUs, considering gaps in knowledge, data, operation, and practice-theory?"*

A research methodology was then proposed to address each gap that would enable the development of reproducible and usable data-driven approaches. Firstly, the IDAIC framework was developed to integrate the concepts, philosophies and methodologies of both the engineering and data science domain which addressed the knowledge gap. Secondly, the IDAIC framework was implemented on an industrial partner site which led to the curation of an industrial AHU dataset and the development of a data assessment decision tree which addressed the data gap. Furthermore, the IDAIC implementation resulted in the novel extension of the APAR ruleset which, along with the findings of the data assessment, led to the discovery of unnoticed inefficiencies which contributed towards addressing the operational gap. Finally, the IDAIC framework was deployed as a tool in the industrial partner site which elicited, and tackled, the practical considerations of data acquisition, resource constraints, cyber security concerns and latency issues that contributed towards closing the practice-theory gap. The following subsection discusses the main findings from each of these research activities in relation to the research questions.

7.2 Conclusions in Respect of Research Questions

RQ 1: *What distinguishes engineering and data science in terms of problem-solving abilities, and how can the strengths of both disciplines be integrated to develop a proactive maintenance solution for AHUs in the industrial setting?*

The critical appraisal of the literature suggested that engineering techniques such as knowledge-based FDD are suited for goal-oriented problem-solving of the known or visible faults, while data science techniques such as exploratory data analysis are suited for exploratory-based problem-solving of the unnoticed or invisible faults. The main finding from the framework development is that the data science approach may be enabled by extending the engineering-based approach to include the activities that are pre-requisite for invisible problem-solving. The addition of a dedicated phase for data assessment, operation assessment and commissioning to the CRISP-DM process model aids in the detection of the issues with AHU data and AHU operation that are acting as a barrier to the proactive data-driven approach. Therefore, effective visible problem-solving which is physically resolved in the act of commissioning enables the value of invisible problem-solving which incorporates the strengths of both the engineering and data science domains.

RQ 2: *How can raw AHU data from the industrial setting be systematically curated into a dataset that enables the development of practical and reproducible data-driven analysis?*

A main finding from the curation of the AHU dataset is that the data from the industrial partner site is difficult to prepare. The reasons include but are not limited to, bespoke data collection methods, bespoke data naming conventions, and heterogeneous sources. This creates a significant challenge to convert the data to a form ready for analysis. Therefore, the prospect of analysing the data is dependent on the skill of the data analyst and the elicitation of the site-specific intricacies that need to be overcome. The IDAIC framework effectively guided the data preparation process in this regard, providing the structure to evoke site-specific knowledge and to incorporate data naming standards in the domain understanding and data contextualisation phases. Furthermore, a key contribution is made to the academic community by providing the curated industrial AHU dataset open source.

To then enable the development of practical and reproducible data-driven analysis, the dataset is assessed to uncover the data issues that need to be resolved. While a novel data assessment decision tree was successfully developed to detect these issues, the lack of action from the industrial partner site to rectify them suggests that further justification is needed to motivate change. While the presence of data issues causes uncertainty in the interpretation of the results of data-driven approaches, and therefore a lack of trust in these techniques, their detection is critical. Understanding the nature of the data issues aids the development of practical, reproducible amendments to data-driven approaches which provide value to decision-makers in the industrial setting.

Therefore, the IDAIC framework enables the systematic curation of an AHU dataset. This, in turn, enables an assessment that contributes to the identification of data issues, making the analysis practical and reproducible through the structured approach taken.

RQ 3: *How can the operation of a real-world AHU in an industrial facility be assessed to identify the sensor, component, and control faults that must be addressed prior to the application of data-driven approaches for the purpose of proactively maintaining the AHU?*

Building on the development of the data assessment decision tree to detect the data issues, or sensor faults, knowledge-based FDD is incorporated in the operation assessment phase to detect component and control faults. The APAR rule-based algorithm was selected as an appropriate means to detect the component and control faults which required novel extension to enable application during the COVID-19 pandemic. The control fault relating to the failure to return to normal operation after the COVID-19 pandemic was a significant energy-consuming fault, which on rectification resulted in approximately €60,000 in annualised heating savings.

Given that the "as is" application of expert rules is not always correct, albeit as a result of a pandemic, the integration of domain expertise to provide contextual understanding is needed to address the operational gap. Without this critical element of domain understanding, the analysis could not have explained the COVID_Operation and the savings could not have been achieved.

RQ 4: *What practical considerations should be taken into account to enable reproducibility in the development of a tool for the purposes of proactively maintaining an AHU in the industrial setting?*

Finally, the analysis was deployed as a tool to uncover practical considerations that contribute to the practice-theory gap. The main findings from the deployment are the difficulties in overcoming data access issues, resource constraints, cyber security concerns and latency issues. The successful development of a tool for proactive maintenance proved the effectiveness of the IDAIC framework for tool development. Furthermore, the UML integration showed the flexibility of the IDAIC framework to work with modelling languages. This further proves the practicality of the framework-based approach to enable software-agnostic deployment that is reproducible.

While the IDAIC framework provided the guidance and structure to overcome these issues, the importance of the skills and knowledge of the tool developers cannot be understated. The strong collaboration between partners on the DENiM EU project was critical in the timely development of the tool during the course of this research.

Table 7-1 Summary of research objectives and contributions

Gap	Contribution	Output
Knowledge	IDAIC Framework - Integration-type adaptation of CRISP-DM to enable visible and invisible problem-solving in a hybrid approach	Journal Article 1
Data	Systematic curation of an industrial AHU dataset & data assessment decision tree	Journal Article 2
Operational	Novel extension to the APAR ruleset	Journal Article 3
Practice-theory	Novel integration of UML and IDAIC for guided deployment of a tool	Journal Article 4

7.3 Research Limitations

Given that the gaps identified in the literature suggested that industry was stuck in reactive mode in relation to AHU maintenance, this research focussed on the development of an approach to proactively maintaining AHUs. This approach was subsequently implemented on a single AHU, which integrated onsite expertise from a single case study site. Therefore, the main limitations of this research may be classified in terms of scope, reliance on domain knowledge, and the industrial-focus.

Firstly, in relation to scope, the IDAIC framework was implemented on only one AHU. This leads to limitations in terms of the breadth of this research. For example, the investigated AHU was a return-air unit, which only operated from the perspective of providing adequate temperature control and adequate ventilation only. Therefore, this research does not provide guidance for other AHU types such as those that provide humidification or heat recovery by a thermal wheel for example.

Secondly, the reliance on domain knowledge is a limitation of this research in terms of scalability. While the integration of domain knowledge is key to address the knowledge gap, it’s a potential source of error when used for the selection of thresholds for fault detection purposes. Additionally, the need for onsite expertise to contextualise raw data and provide undocumented site information, such as COVID_Operation, is another limitation if this knowledge is not readily available.

Lastly, this research focussed on the nuances of the industrial setting. The applicability of this research in the commercial or residential setting is not tested. These building types may experience different issues in comparison to a heavily regulated medical device manufacturing facility. The less regulated nature of these buildings may potentially allow for additional sensors to be installed, which would lead to further analytical capabilities given the increased availability of data.

7.4 Recommendations for Future Research

This research presents a framework that enables the integration of domain expertise and data-driven analysis to proactively maintain AHUs. As discussed in the previous subsection, this research contains limitations which may be addressed in future work. There are two main potential research avenues that may emanate from this research, scope expansion and further technical analysis, which contain multiple research tasks that may be completed over both the medium-term (within one year) and long-term (longer than one year) as shown in Figure 7-1. A suggested progression between the research tasks is also provided as a guide.

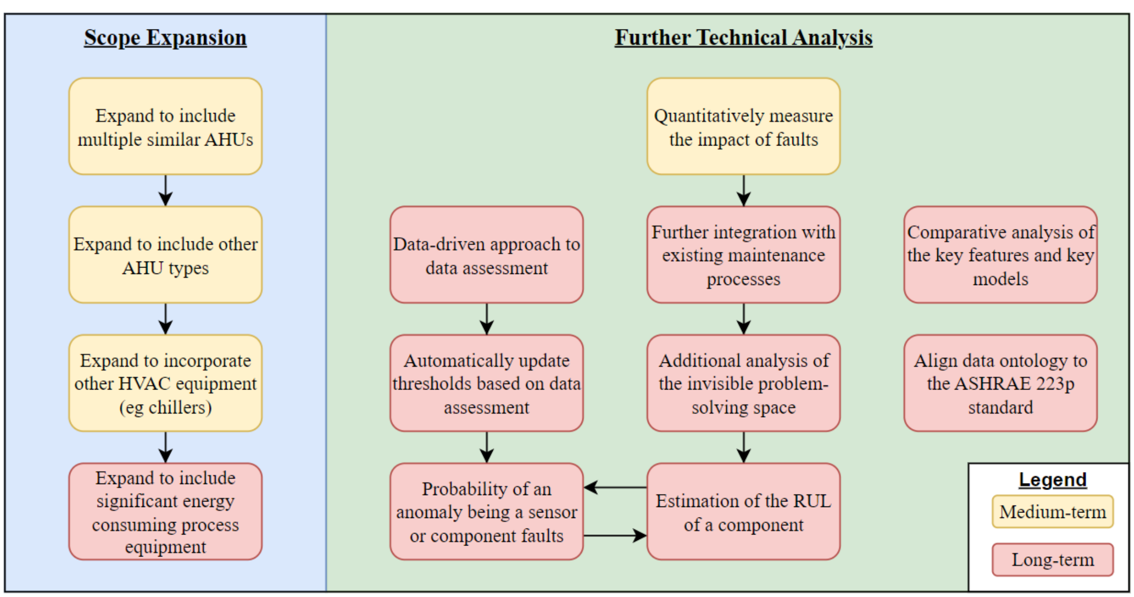


Figure 7-1 Areas for further research

Firstly, this research has the potential to be expanded to other energy-consuming assets. The main tasks include:

- **Expansion to other HVAC equipment -** The IDAIC framework was developed for the specific purpose of maintaining AHUs, however, an opportunity exists to extend the implementation of the IDAIC framework to other HVAC equipment in the medium-term. For example, the AHU selected for this research is one of four similar AHUs on the case study site. Analysis of these similar AHUs would enable an accurate comparison of their operational efficiency. Furthermore, other types of AHUs may be investigated, along with other HVAC components such as chillers for example.
- **Expansion to process equipment -** Beyond the HVAC system, the production equipment in industrial sites is another area of future research. Given the bespoke nature of production equipment, there is likely to be a lack of shallow knowledge rules readily available in the literature for application. The IDAIC framework may then need an extension to guide the development of shallow knowledge rules for this purpose, which is a longer term research area.

Secondly, the following research areas could potentially add value from a more in-depth technical analysis perspective:

- **Data-driven approach to data assessment phase –** Given the siloed nature of the dataset curated in this research, there were many issues identified using the integrated expert knowledge and statistical property decision tree. However, as the Industry 5.0 paradigm matures and industrial sites become more digitalised, the data issues are likely to be rectified in new systems. Therefore, the deviations of the data from the true physical property they measure may be less frequent and more difficult to detect. A data-driven approach may be applicable in these circumstances to detect subtle deviations from normality and notify decision-makers of their occurrence.
- **Threshold selection based on the data assessment–** In the data assessment phase, the concept of analysing the inconsistencies over a period of time was

introduced. The temperature differences between similar sensors were analysed over a short one-hour period of only four data points to uncover multiple sensor fault types. Similarly, APAR and the novel extension for COVID_Operation also aim to detect inconsistencies between sensor values and control signals, but at an instantaneous point in time. However, the consistency is only analysed when the valves are fully shut. This may be extended to analyse the inconsistency at an open position, similar to the data exploration conducted in the framework implementation chapter. This work may then be used to inform the selection of appropriate thresholds for expert rule application, or perhaps automatically update them.

- **Quantitatively measure the impact of faults –** In the framework implementation chapter, an annualised saving of €60,000 was calculated based on assumptions made about the AHUs operation. In future work, the estimated savings may be more accurately obtained by including data from the chiller and the boiler. This would enable a more accurate calculation of the chiller and boiler efficiency at the time a fault occurred. Furthermore, the impact of the fault may then be measured in relation to the overall energy consumption increase that has occurred while the fault was active. This would enable accurate quantification of the impact of the fault to aid the prioritisation of maintenance actions by the facilities team of an industrial site.
- **Further integration with existing maintenance processes –** In the development of the IDAIC framework, the lack of integration with existing strategies was identified as a shortcoming of CRISP-DM for the purpose of proactively maintaining an AHU. This shortcoming was addressed by incorporating a commissioning phase which aimed to establish a fault-free baseline operation to enable invisible problem-solving. While most data-driven approaches generally require a fault-free baseline, the supervised approaches also require labelled data. Therefore, IDAIC may be further extended into the maintenance processes on an industrial site to capture the maintenance activity that has taken place. This will require close long-term collaboration with the maintenance team to systematically capture the activity that has taken place. This research may involve the

standardisation of documenting maintenance activity, such that it may be easily integrated as data labels into the invisible problem analysis of the IDAIC framework.

- **Additional analysis of the invisible problem-solving space –** The visible problems that remain on the AHU in the industrial partner site created difficulty in the application of the final two phases of the IDAIC framework. While a fault-free baseline could not be obtained, the COVID_Operation provided a means of continuing the invisible problem-solving activity to show how the latter phases of the IDAIC framework work. Further guidance on the algorithm selection and results exploration was also provided in the tool deployment. However, in future work on a fault-free AHU, the final two phases may be more thoroughly applied to compare the performance of data-driven approaches proposed in the literature for the detection of subtle deviations from normality. This research would require a strong commitment from the industrial site to dedicate resources for a full year of operation to ensure that the identified data issues and operational issues with the AHU are promptly addressed.
- **Further analysis of the fault nature –** The bottom two research tasks, estimating the probability an anomaly is a sensor or component fault and estimating the remaining useful life (RUL) of a component, would both contribute to an increased understanding of the nature of a faults' occurrence. These research activities are closely linked and follow a natural progression from automatically updating thresholds and further analysis of invisible problems. These tasks are distinct from the previously suggested tasks in that the focus is on developing a quantitative metric for decision support. The nature of these inconsistencies, or the degradation observed, could be further analysed to detect the likelihood that the inconsistency is a sensor or component fault by further leveraging the historical data in a data-driven approach. This would benefit a commissioning engineer by estimating the probability a fault is caused by a sensor issue or component issue, with further guidance provided to an energy manager in relation to the estimation of the RUL of a component to aid the scheduling of maintenance activity.

- **Comparative analysis of the key features and key models –** As industry progresses further along the path to digitalisation-based approaches, the possibility of model comparison may arise. Research could aim to find the minimum set of sensors required to apply data-driven FDD. Furthermore, the specific sensors required to identify specific faults could be investigated. This would provide industrial practitioners with a means of identifying the sensors required to detect the faults of most interest.
- **Aligning to ASHRAE 223p standard –** The Project Haystack initiative was leveraged in this research to contextualise the raw data from the case study site. However, efforts are currently ongoing to develop the ASHRAE 223p standard, which aims to integrate Haystack tagging and Brick data modelling concepts. Future work may investigate the most effective means to align to this standard, which would considerably increase the applicability of FDD methods for proactively maintaining AHUs.

8 References

[1] IPCC, “SYNTHESIS REPORT OF THE IPCC SIXTH ASSESSMENT REPORT (AR6) IPCC,” 2023.

[2] U. Nations, “Transforming our world: the 2030 agenda for sustainable development,” *Rep. No. A/RES/70/1*.

[3] I. G. of S. appointed by the Secretary-General, “Global Sustainable Development Report 2019: The Future is Now – Science for Achieving Sustainable Development,” New York, 2019.

[4] European Commission, “An industrial strategy for Europe,” 2020.

[5] E. I. Bank, “The 2021/2022 digitalisation in Europe report : The pandemic has made the digital transformation an integral part of European society,” pp. 1–6, 2022.

[6] European Commission, “A New Industrial Strategy for Europe,” 2020.

[7] International Energy Agency, “International Energy Agency (IEA) World Energy Outlook 2022,” *Https://Www.Iea.Org/Reports/World-Energy-Outlook-2022/Executive-Summary*, p. 524, 2022.

[8] W. Nelson and C. Culp, “Machine Learning Methods for Automated Fault Detection and Diagnostics in Building Systems-A Review,” *energies*, 2022.

[9] K. W. Roth, D. Westphalen, M. Y. Feng, P. Llana, and L. Quartararo, “Energy Impact of Commercial Building Controls and Performance Diagnostics: Market Characterization, Energy Impact of Building Faults and Energy Savings Potential,” 2005.

[10] E. Mills, “Building commissioning: A golden opportunity for reducing energy costs and greenhouse gas emissions in the United States,” *Energy Effic.*, vol. 4, no. 2, pp. 145–173, 2011.

[11] M. Achouch *et al.*, “Predictive Maintenance and Fault Monitoring Enabled by Machine Learning: Experimental Analysis of a TA-48 Multistage Centrifugal Plant Compressor,” *Appl. Sci.*, vol. 13, no. 3, 2023.

[12] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, “Prognostics and health management design for rotary machinery systems — Reviews , methodology and applications,” *Mech. Syst. Signal Process.*, vol. 42, no. 1–2, pp. 314–334, 2014.

[13] J. Granderson, G. Lin, A. Harding, P. Im, and Y. Chen, “Building fault detection data to aid diagnostic algorithm creation and performance testing,” *Sci. Data*, pp. 1–14, 2020.

[14] J. Granderson, G. Lin, R. Singla, E. Mayhorn, P. Ehrlich, and D. Vrabie, “Commercial Fault Detection and Diagnostics Tools: What They Offer, How They Differ, and What’s Still Needed,” *Lawrence Berkeley Natl. Lab. Recent Work*, pp. 1–12, 2018.

[15] J. Granderson and G. Lin, "Inventory of Data Sets for AFDD Evaluation, Building Technology and Urban Systems Division," 2019.

[16] D. E. Claridge *et al.*, "Is commissioning once enough?," *Energy Eng. J. Assoc. Energy Eng.*, vol. 101, no. 4, pp. 7–19, 2004.

[17] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, "A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis," *Energy Built Environ.*, vol. 1, no. 2, pp. 149–164, 2020.

[18] Y. Zhao, T. Li, X. Zhang, and C. Zhang, "Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future," *Renew. Sustain. Energy Rev.*, vol. 109, no. April, pp. 85–101, 2019.

[19] J. Chen, L. Zhang, Y. Li, Y. Shi, X. Gao, and Y. Hu, "A review of computing-based automated fault detection and diagnosis of heating, ventilation and air conditioning systems," *Renew. Sustain. Energy Rev.*, vol. 161, no. July 2021, p. 112395, 2022.

[20] M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine Learning for industrial applications : A comprehensive literature review," *Expert Syst. Appl.*, vol. 175, no. March 2020, p. 114820, 2021.

[21] K. Lepenioti, A. Bousdekis, D. Apostolou, and G. Mentzas, "Prescriptive analytics: Literature review and research challenges," *Int. J. Inf. Manage.*, vol. 50, no. October 2018, pp. 57–70, 2020.

[22] European Commission, "Annex to the Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions—The European Green Deal. Roadmap—Key Actions," 2019.

[23] Government of Ireland, "Climate Action Plan 2023," 2023.

[24] International Energy Agency, "7th Annual Global Conference on Energy Efficiency: The value of urgent action on energy efficiency," p. 12, 2022.

[25] "Local Law 97 - Sustainable Buildings." [Online]. Available: https://www.nyc.gov/site/sustainablebuildings/ll97/local-law-97.page. [Accessed: 27-Apr-2023].

[26] Senate and House of Representatives of the United States of America, "Inflation Reduction Act of 2022, Pub. L. No. 117–169, 136 Stat. 1818," pp. 1–274, 2022.

[27] EU, "DIRECTIVE (EU) 2018/844 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," *Off. J. Eur. Union*, vol. 276 LNCS, no. May, pp. 98–107, 2018.

[28] B. Precious, "Energy efficiency in commercial buildings," 2022.

[29] J. A. Garza-Reyes, "Lean and green-a systematic review of the state of the art literature," *J. Clean. Prod.*, vol. 102, pp. 18–29, 2015.

[30] S. Prasad and S. K. Sharma, “Lean and Green Manufacturing: Concept and its Implementation in Operations Management,” *Int. J. Adv. Mech. Eng.*, vol. 4, no. 5, pp. 2250–3234, 2014.

[31] A. Fercoq, S. Lamouri, V. Carbone, A. Lelièvre, and A. A. Lemieux, *Combining lean and green in manufacturing: A model of waste management*, vol. 46, no. 9. IFAC, 2013.

[32] J. Hill, A. J. Thomas, R. K. Mason-Jones, and S. El-Kateb, “The implementation of a Lean Six Sigma framework to enhance operational performance in an MRO facility,” *Prod. Manuf. Res.*, vol. 6, no. 1, pp. 26–48, 2018.

[33] T. Zonta *et al.*, “Predictive maintenance in the Industry 4 . 0 : A systematic literature review,” *Comput. Ind. Eng.*, vol. 150, no. August, p. 106889, 2020.

[34] M. Núñez-Merino, J. M. Maqueira-Marín, J. Moyano-Fuentes, and P. J. Martínez-Jurado, “Information and digital technologies of Industry 4.0 and Lean supply chain management: a systematic literature review,” *Int. J. Prod. Res.*, vol. 58, no. 16, pp. 5034–5061, 2020.

[35] A. Bueno, M. Godinho Filho, and A. G. Frank, “Smart production planning and control in the Industry 4.0 context: A systematic literature review,” *Comput. Ind. Eng.*, vol. 149, no. August, p. 106774, 2020.

[36] P. Rosa, C. Sassanelli, A. Urbinati, and D. Chiaroni, “Assessing relations between Circular Economy and Industry 4 . 0 : a systematic literature review,” vol. 7543, 2020.

[37] G. of Canada, “Pan-Canadian Artificial Intelligence Strategy,” 2017. [Online]. Available: https://ised-isde.canada.ca/site/ai-strategy/en. [Accessed: 13-Mar-2023].

[38] “State Council Issued Notice of the New Generation Artificial Intelligence Development Plan.” [Online]. Available: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. [Accessed: 13-Mar-2023].

[39] European Commission, “Coordinated Plan on AI,” *Commun. EC*, 2018.

[40] OECD, “AI FOR HUMANITY - NATIONAL STRATEGY IN ARTIFICIAL INTELLIGENCE.” [Online]. Available: https://stip.oecd.org/stip/interactive-dashboards/policy-initiatives/2021%2Fdata%2FpolicyInitiatives%2F24070. [Accessed: 13-Mar-2023].

[41] German Federal Government, “Artificial Intelligence Strategy of Germany Federal Government,” *Ki startegie AI Made Ger.*, no. December, p. 31, 2018.

[42] T. D. Government, “National Strategy for Artificial Intelligence,” no. March, 2019.

[43] Cabinet Office Japan, “AI Strategy 2022,” vol. 2022, 2022.

[44] Ministry of Science and ICT Artificial Intelligence Policy Division, “National Strategy for Artificial intelligence,” *Gov. Repub. Korea*, p. 114, 2019.

[45] P. of the M. of E. A. and E. Enterprises, “Artificial Intelligence 4.0 programme,” 2022.

[46] National Artificial Intelligence Research Resource Task Force, “Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource,” no. January, 2023.

[47] Government of Ireland, “AI - here for good: A national artifical intelligence strategy for Ireland,” *Dep. Enterp. Trade Employ.*, 2021.

[48] DCMS, “National AI Strategy,” *Off. AI*, p. 35, 2021.

[49] I. Government, “Strategic Programme on Artificial Intelligence,” 2022.

[50] E. Commission, Directorate-General for Research and Innovation, A. Renda, S. Schwaag Serger, D. Tataj, and E. Al., “Industry 5.0, a transformative vision for Europe : governing systemic transformations towards a sustainable industry,” 2022.

[51] E. Commission and D.-G. for R. and Innovation, *Industry 5.0 : human-centric, sustainable and resilient*. Publications Office, 2021.

[52] “Industry 5.0.” [Online]. Available: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en#how-to-make-it-happen. [Accessed: 13-Jul-2023].

[53] R. Clancy, K. Bruton, D. O’Sullivan, and D. Keogh, “Industry 4.0 driven statistical analysis of investment casting process demonstrates the value of digitalisation,” *Procedia Comput. Sci.*, vol. 200, pp. 284–297, 2022.

[54] U.S. EIA, “Annual Energy Outlook 2022,” 2022.

[55] Global e-Sustainability Initiative, “#SMARTer2030-ICT Solutions for 21st Century Challenges,” 2015.

[56] IRENA, “Global Energy Transformation: A Roadmap to 2050,” 2018.

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

[58] CIM, “THE ENERGY BLIND SPOTS,” 2022.

[59] P. O. Donovan, K. Leahy, K. Bruton, and D. T. J. O. Sullivan, “An industrial big data pipeline for data - driven analytics maintenance applications in large - scale smart manufacturing facilities,” *J. Big Data*, pp. 1–26, 2015.

[60] B. S. Dhillon, *Maintainability, Maintenance , and Reliability for Engineers*. CRC Press, 2006.

[61] D. E. Claridge, C. H. Culp, M. Liu, S. Deng, W. D. Turner, and J. S. Haberl, “Campus-wide continuous commissioningSM of university buildings,” *Proc. ACEEE Summer Study Energy Effic. Build.*, vol. 3, pp. 3101–3111, 2000.

[62] J. Schein, S. T. Bushby, N. S. Castro, and J. M. House, “Results from Field Testing

of Air Handling Unit and Variable Air Volume Box Fault Detection Tools," 2003.

[63] G. Bianco, F. Delfino, G. Ferro, M. Robba, and M. Rossi, "A hierarchical Building Management System for temperature's optimal control and electric vehicles' integration," *Energy Convers. Manag. X*, vol. 17, no. November 2022, p. 100339, 2023.

[64] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, and F. S. García, "A review of strategies for building energy management system : Model predictive control , demand side management , optimization , and fault detect & diagnosis," vol. 33, no. July 2020, 2021.

[65] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustain. Cities Soc.*, vol. 48, no. April, p. 101533, 2019.

[66] "ASHRAE Guideline 36."

[67] "OpenBlue | Johnson Controls." [Online]. Available: https://www.johnsoncontrols.com/openblue. [Accessed: 03-May-2023].

[68] "The PEAK Platform | Building Data Analytics Software by CIM." [Online]. Available: https://www.cim.io/the-peak-platform/overview. [Accessed: 03-May-2023].

[69] "Fault-Diagnosis Applications: Model-Based Condition Monitoring: Actuators ... - Rolf Isermann - Google Books." [Online]. Available: https://books.google.ie/books?hl=en&lr=&id=v2EOBm7bTJgC&oi=fnd&pg=PR3&ots=VG5bOO2Mw1&sig=byTU3R2ycoMRgI7crkeMX21OSw0&redir_esc=y#v=onepage&q&f=false. [Accessed: 07-Sep-2022].

[70] J. Lee, H. Davari, J. Singh, and V. Pandhare, "Industrial Artificial Intelligence for industry 4.0-based manufacturing systems," *Manuf. Lett.*, vol. 18, pp. 20–23, 2018.

[71] L. M. Pintelon and L. F. Gelders, "Maintenance management decision making," *Eur. J. Oper. Res.*, vol. 58, no. 3, pp. 301–317, 1992.

[72] K. Bruton, P. Raftery, and B. Kennedy, "Review of automated fault detection and diagnostic tools in air handling units," pp. 335–351, 2014.

[73] AiCARR, ASHRAE, and CIBSE, *Commissioning Definitions and Terminology for the Building Industry : A Common Overview*. 2018.

[74] D. Bumblauskas, D. Gemmill, A. Igou, and J. Anzengruber, "Smart Maintenance Decision Support Systems (SMDSS) based on corporate big data analytics," *Expert Syst. Appl.*, vol. 90, pp. 303–317, 2017.

[75] W. Tiddens, J. Braaksma, and T. Tinga, "Decision Framework for Predictive Maintenance Method Selection," *Appl. Sci.*, vol. 13, no. 3, 2023.

[76] B. S. Dhillon, *ENGINEERING MAINTENANCE*. .

[77] A. Prajapati, J. Bechtel, and S. Ganesan, "Condition based maintenance: A survey," *J. Qual. Maint. Eng.*, vol. 18, no. 4, pp. 384–400, 2012.

[78] E. I. Basri, I. H. A. Razak, H. Ab-Samat, and S. Kamaruddin, "Preventive maintenance (PM) planning: A review," *J. Qual. Maint. Eng.*, vol. 23, no. 2, pp. 114–143, 2017.

[79] N. Harris, T. Shealy, H. Kramer, J. Granderson, and G. Reichard, "A framework for monitoring-based commissioning: Identifying variables that act as barriers and enablers to the process," *Energy Build.*, vol. 168, pp. 331–346, 2018.

[80] J. Lee, *Industrial AI: Applications with Sustainable Performance*. Springer, 2020.

[81] F. Ansari, R. Glawar, and T. Nemeth, "PriMa: a prescriptive maintenance model for cyber-physical production systems," *Int. J. Comput. Integr. Manuf.*, vol. 32, no. 4–5, pp. 482–503, 2019.

[82] S. Katipamula and M. R. Brambley, "Methods for Fault Detection , Diagnostics , and Prognostics for Building Systems — A Review , Part I," vol. 11, no. 1, pp. 3–26, 2005.

[83] S. P. Melgaard, K. H. Andersen, A. Marszal-Pomianowska, R. L. Jensen, and P. K. Heiselberg, "Fault Detection and Diagnosis Encyclopedia for Building Systems: A Systematic Review," *Energies*, vol. 15, no. 12, p. 4366, 2022.

[84] "Home : Oxford English Dictionary." [Online]. Available: https://www.oed.com/. [Accessed: 02-May-2023].

[85] C. Shearer *et al.*, "The CRISP-DM model: The New Blueprint for Data Mining," *J. Data Warehous.*, vol. 5, no. 4, pp. 13–22, 2000.

[86] D. Hand, H. Mannila, and P. Smyth, *Principles of data mining*, vol. 30, no. 7. 2001.

[87] F. Martínez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 4347, no. c, pp. 1–1, 2019.

[88] R. Wirth and J. Hipp, "CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.

[89] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *Adv. Knowl. Discov. data Min.*, vol. 17, no. 3, pp. 37–54, 1996.

[90] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *Knowl. Eng. Rev.*, vol. 25, no. 2, pp. 137–166, 2010.

[91] "ASUM-DM Teaser." [Online]. Available: http://gforge.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html. [Accessed: 28-Mar-2022].

[92] S. Moyle and A. Jorge, "RAMSYS - A methodology for supporting rapid remote collaborative data mining projects," *Proc. ECML/PKDD'01 Work. Integr. Asp.*

Data Mining, Decis. Support Meta-Learning, no. 1, pp. 20–31, 2001.

[93] F. Martínez-Plumed *et al.*, “Context Aware Standard Process for Data Mining,” *Nature*, 2017. [Online]. Available: https://arxiv.org/abs/1709.09003.

[94] V. Plotnikova, M. Dumas, and F. Milani, “Adaptations of data mining methodologies: A systematic literature review,” *PeerJ Comput. Sci.*, vol. 6, pp. 1–43, 2020.

[95] J. W. Tukey, *Exploratory Data Analysis*. 1977.

[96] A. Azevedo and M. F. Santos, “KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW,” *IADIS Eur. Conf. Data Min.*, pp. 182–185, 2008.

[97] T. Wuest *et al.*, “Machine learning in manufacturing : advantages , challenges , and applications,” *Prod. Manuf. Res.*, vol. 3277, pp. 1–23, 2016.

[98] J. Lenz, T. Wuest, and E. Westkämper, “Holistic approach to machine tool data analytics,” *J. Manuf. Syst.*, vol. 48, pp. 180–191, 2018.

[99] “About Our Center — Site.” [Online]. Available: http://www.imscenter.net/. [Accessed: 07-Apr-2023].

[100] “Key enabling technologies.” [Online]. Available: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/key-enabling-technologies_en. [Accessed: 25-Apr-2023].

[101] J. J. Ortega-Gras, M. V. Bueno-Delgado, G. Cañavate-Cruzado, and J. Garrido-Lova, “Twin transition through the implementation of industry 4.0 technologies: Desk-research analysis and practical use cases in europe,” *Sustain.*, vol. 13, no. 24, 2021.

[102] A. L. Samuel, “Some Studies in Machine Learning,” *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959.

[103] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” vol. 349, no. 6245, 2015.

[104] H. Wang, C. Ma, and L. Zhou, “A brief review of machine learning and its application,” *Proc. - 2009 Int. Conf. Inf. Eng. Comput. Sci. ICIECS 2009*, pp. 9–12, 2009.

[105] I. Matetić, I. Štajduhar, I. Wolf, and S. Ljubic, “A Review of Data-Driven Approaches and Techniques for Fault Detection and Diagnosis in HVAC Systems,” *Sensors*, vol. 23, no. 1, pp. 1–37, 2023.

[106] Z. Chen *et al.*, “A review of data-driven fault detection and diagnostics for building HVAC systems,” *Appl. Energy*, vol. 339, no. March, p. 121030, 2023.

[107] W. Tun, J. K. W. Wong, and S. H. Ling, “Hybrid random forest and support vector machine modeling for hvac fault detection and diagnosis,” *Sensors*, vol. 21, no. 24, pp. 1–15, 2021.

[108] F. Cheng, W. Cai, X. Zhang, H. Liao, and C. Cui, “Fault detection and diagnosis for Air Handling Unit based on multiscale convolutional neural networks,” *Energy*

Build., vol. 236, p. 110795, 2021.

[109] M. Elnour and N. Meskin, “Novel Actuator Fault Diagnosis Framework for Multizone HVAC Systems Using 2-D Convolutional Neural Networks,” *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 1985–1996, 2021.

[110] A. Rosato, F. Guarino, S. Sibilio, E. Entchev, M. Masullo, and L. Maffei, “Healthy and faulty experimental performance of a typical HVAC system under italian climatic conditions: Artificial neural network-based model and fault impact assessment,” *Energies*, vol. 14, no. 17, 2021.

[111] W. Yun, W. Hong, and H. Seo, “A data-driven fault detection and diagnosis scheme for air handling units in building HVAC systems considering undefined states,” *J. Build. Eng.*, vol. 35, no. December 2020, p. 102111, 2021.

[112] H. Liao, W. Cai, F. Cheng, S. Dubey, and P. B. Rajesh, “An online data-driven fault diagnosis method for air handling units by rule and convolutional neural networks,” *Sensors*, vol. 21, no. 13, 2021.

[113] Z. Liu, Z. Huang, J. Wang, C. Yue, and S. Yoon, “A novel fault diagnosis and self-calibration method for air-handling units using Bayesian Inference and virtual sensing,” *Energy Build.*, vol. 250, p. 111293, 2021.

[114] S. Taheri, A. Ahmadi, B. Mohammadi-Ivatloo, and S. Asadi, “Fault detection diagnostic for HVAC systems via deep learning algorithms,” *Energy Build.*, vol. 250, p. 111275, 2021.

[115] B. Wu, W. Cai, H. Chen, and X. Zhang, “A hybrid data-driven simultaneous fault diagnosis model for air handling units,” *Energy Build.*, vol. 245, p. 111069, 2021.

[116] A. Taal and L. Itard, “Fault detection and diagnosis for indoor air quality in DCV systems: Application of 4S3F method and effects of DBN probabilities,” *Build. Environ.*, vol. 174, no. January, p. 106632, 2020.

[117] M. Parzinger, L. Hanfstaengl, F. Sigg, U. Spindler, U. Wellisch, and M. Wirnsberger, “Residual analysis of predictive modelling data for automated fault detection in building’s heating, ventilation and air conditioning systems,” *Sustain.*, vol. 12, no. 17, 2020.

[118] C. Zhong, K. Yan, Y. Dai, N. Jin, and B. Lou, “Energy Efficiency Solutions for Buildings: Automated Fault Diagnosis of Air Handling Units Using Generative Adversarial Networks,” *Energies*, vol. 12, no. 3, pp. 1–11, 2019.

[119] D. Li, Y. Zhou, G. Hu, and C. J. Spanos, “Handling Incomplete Sensor Measurements in Fault Detection and Diagnosis for Building HVAC Systems,” *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 833–846, 2019.

[120] Z. Ghahramani, “Unsupervised Learning,” *Adv. Lect. Mach. Learn.*, vol. 55, pp. 80–112, 2004.

[121] A. A. Markus, B. W. Hobson, H. B. Gunay, and S. Bucking, “A framework for a multi-source , data-driven building energy management toolkit,” *Energy Build.*, vol. 250, p. 111255, 2021.

[122] L. Burgas, J. Colomer, J. Melendez, F. I. Gamero, and S. Herraiz, “Integrated Unfold-PCA Monitoring Application for Smart Buildings : An AHU Application Example,” 2021.

[123] H. Zhang, C. Li, D. Li, Y. Zhang, and W. Peng, “Fault detection and diagnosis of the air handling unit via an enhanced kernel slow feature analysis approach considering the time-wise and batch-wise dynamics,” *Energy Build.*, vol. 253, p. 111467, 2021.

[124] H. B. Gunay and Z. Shi, “Cluster analysis-based anomaly detection in building automation systems,” *Energy Build.*, vol. 228, p. 110445, 2020.

[125] K. Yan and W. Shen, “Unsupervised Learning for Fault Detection and Diagnosis of Air Handling Units,” *Energy Build.*, no. January 2022, 2019.

[126] H. Hassanpour, P. Mhaskar, J. M. House, and T. I. Salsbury, “A hybrid modeling approach integrating first-principles knowledge with statistical methods for fault detection in HVAC systems,” *Comput. Chem. Eng.*, vol. 142, p. 107022, 2020.

[127] H. Zhu, W. Yang, S. Li, and A. Pang, “An Effective Fault Detection Method for HVAC Systems Using the LSTM-SVDD Algorithm,” *Buildings*, vol. 12, no. 2, p. 246, 2022.

[128] C. Fan, Y. Liu, X. Liu, Y. Sun, and J. Wang, “A study on semi-supervised learning in enhancing performance of AHU unseen fault detection with limited labeled data,” *Sustain. Cities Soc.*, vol. 70, no. December 2020, p. 102874, 2021.

[129] C. Fan, X. Liu, P. Xue, and J. Wang, “Statistical characterization of semi-supervised neural networks for fault detection and diagnosis of air handling units,” *Energy Build.*, vol. 234, p. 110733, 2021.

[130] M. S. Piscitelli, D. M. Mazzarelli, and A. Capozzoli, “Enhancing operational performance of AHUs through an advanced fault detection and diagnosis process based on temporal association and decision rules,” *Energy Build.*, vol. 226, p. 110369, 2020.

[131] V. Martinez-viol, M. Delgado-prieto, E. M. Urbano, and L. Romeral, “Support vector machine based novelty detection and FDD framework applied to building AHU systems,” pp. 1749–1754, 2020.

[132] M. S. Mirnaghi and F. Haghighat, “Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review,” *Energy Build.*, vol. 229, p. 110492, 2020.

[133] J. Huang *et al.*, “Real vs. simulated: Questions on the capability of simulated datasets on building fault detection for energy efficiency from a data-driven perspective,” *Energy Build.*, vol. 259, p. 111872, 2022.

[134] V. Singh, J. Mathur, and A. Bhatia, “A Comprehensive Review: Fault Detection, Diagnostics, Prognostics, and Fault Modelling in HVAC Systems,” *Int. J. Refrig.*, vol. 144, no. July, pp. 283–295, 2022.

[135] L. Da Xu, E. L. Xu, and L. Li, “Industry 4.0: state of the art and future trends,” *Int.*

J. Prod. Res., vol. 56, no. 8, pp. 2941–2962, Apr. 2018.

[136] K. Bruton, D. Coakley, P. Raftery, D. O. Cusack, M. M. Keane, and D. T. J. O'Sullivan, "Comparative analysis of the AHU InFO fault detection and diagnostic expert tool for AHUs with APAR," *Energy Effic.*, vol. 8, no. 2, pp. 299–322, 2015.

[137] "RP-1312 -- Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Units | ASHRAE Store." [Online]. Available: https://www.techstreet.com/standards/rp-1312-tools-for-evaluating-fault-detection-and-diagnostic-methods-for-air-handling-units?product_id=1833299. [Accessed: 25-Mar-2021].

[138] G. Lin, H. Kramer, and J. Granderson, "Building fault detection and diagnostics: Achieved savings, and methods to evaluate algorithm performance," *Build. Environ.*, vol. 168, no. June 2019, p. 106505, 2020.

[139] Y. Yu, D. Woradechjumroen, and D. Yu, "A review of fault detection and diagnosis methodologies on air-handling units," *Energy Build.*, vol. 82, pp. 550–562, 2014.

[140] C. M. Fan and Y. P. Lu, "A bayesian framework to integrate knowledge-based and data-driven inference tools for reliable yield diagnoses," *Proc. - Winter Simul. Conf.*, pp. 2323–2329, 2008.

[141] J. M. House, H. Vaezi-Nejad, and J. M. Whitcomb, "An expert rule set for fault detection in air-handling units / Discussion," 2001.

[142] L. Song and G. Wang, "Using a hybrid method to construct a computational efficient cooling coil model for an automated single-duct variable air volume system fault detection and diagnosis," *Energy Build.*, vol. 92, pp. 363–373, 2015.

[143] H. Wang, Y. Chen, C. W. H. Chan, J. Qin, and J. Wang, "Online model-based fault detection and diagnosis strategy for VAV air handling units," *Energy Build.*, vol. 55, pp. 252–263, 2012.

[144] S. Deshmukh, S. Samouhos, L. Glicksman, and L. Norford, "Fault detection in commercial building VAV AHU : A case study of an academic building," *Energy Build.*, vol. 201, pp. 163–173, 2019.

[145] W. Kim and S. Katipamula, "A review of fault detection and diagnostics methods for building systems," *Sci. Technol. Built Environ.*, vol. 24, no. 1, pp. 3–21, 2017.

[146] K. Tidriri, N. Chatti, S. Verron, and T. Tiplica, "Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges," *Annu. Rev. Control*, vol. 42, pp. 63–81, 2016.

[147] S. Ahamed, R. Zmeureanu, N. Cortrufo, and J. Candanedo, "Gray-box virtual sensor of the supply air temperature of air handling units Gray-box virtual sensor of the supply air temperature of air handling units," *Sci. Technol. Built Environ.*, vol. 0, no. 0, pp. 1–12, 2020.

[148] A. Brem, S. K. Bruton, P. O. Sullivan, and A. Cashman, "Advancing Industrial Participation in Demand Response within National Electricity Grids through Systematic Asset Selection , Risk Assessment and Modelling Approaches," no.

August, 2022.

[149] “Are all approaches equal: the difference between technique, tool and method.” [Online]. Available: https://greensam.eu/technique-tool-method/. [Accessed: 27-Apr-2023].

[150] J. Lee, J. Singh, M. Azamfar, and K. Sun, “Industrial AI: A Systematic Framework for AI in Industrial Applications,” *Zhongguo Jixie Gongcheng/China Mech. Eng.*, vol. 31, no. 1, pp. 37–48, 2020.

[151] S. U. Jan, Y. D. Lee, and I. S. Koo, “A distributed sensor-fault detection and diagnosis framework using machine learning,” *Inf. Sci. (Ny).*, vol. 547, pp. 777–796, 2021.

[152] A. Dogan and D. Birant, “Machine learning and data mining in manufacturing,” *Expert Syst. Appl.*, vol. 166, no. February 2019, p. 114060, 2021.

[153] D. Dey and B. Dong, “A probabilistic approach to diagnose faults of air handling units in buildings,” *Energy Build.*, vol. 130, pp. 177–187, 2016.

[154] S. Li and J. Wen, “Application of pattern matching method for detecting faults in air handling unit system,” *Autom. Constr.*, vol. 43, pp. 49–58, 2014.

[155] W. Y. Lee, J. M. House, and N. H. Kyong, “Subsystem level fault diagnosis of a building’s air-handling unit using general regression neural networks,” *Appl. Energy*, vol. 77, no. 2, pp. 153–170, 2004.

[156] I. Errandonea, S. Beltrán, and S. Arrizabalaga, “Digital Twin for maintenance: A literature review,” *Comput. Ind.*, vol. 123, 2020.

[157] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, “Digital Twin in manufacturing: A categorical literature review and classification,” *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, 2018.

[158] A. A. Markus, B. W. Hobson, H. B. Gunay, and S. Bucking, “Does a knowledge gap contribute to the performance gap? Interviews with building operators to identify how data-driven insights are interpreted,” *Energy Build.*, vol. 268, p. 112238, 2022.

[159] X. Xie, J. Merino, N. Moretti, P. Pauwels, J. Y. Chang, and A. Parlikad, “Digital twin enabled fault detection and diagnosis process for building HVAC systems,” *Autom. Constr.*, vol. 146, no. June 2022, p. 104695, 2023.

[160] CIBSE, *Building control systems*. Oxford: Butterworth-Heinemann, 2000.

[161] P. Chapman *et al.*, “Crisp-Dm,” *SPSS inc*, vol. 78, pp. 1–78, 2000.

[162] D. C. Corrales, A. Ledezma, and J. C. Corrales, “A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal,” no. November, 2015.

[163] O. Almutiry, G. Wills, A. Alwabel, R. Crowder, and R. Walters, “Toward a framework for data quality in cloud-based health information system,” *Int. Conf. Inf. Soc. i-Society 2013*, no. August 2019, pp. 153–157, 2013.

[164] M. Ahern, D. T. J. O. Sullivan, and K. Bruton, "Development of a Framework to Aid the Transition from Reactive to Proactive Maintenance Approaches to Enable Energy Reduction," *Appl. Sci.*, 2022.

[165] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Sci. J.*, vol. 14, pp. 1–10, 2015.

[166] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, 2009.

[167] R. Clancy, D. O'sullivan, and K. Bruton, "Data-driven quality improvement approach to reducing waste in manufacturing."

[168] Z. Shi and W. O. Brien, "Development and implementation of automated fault detection and diagnostics for building systems : A review," *Autom. Constr.*, vol. 104, no. April, pp. 215–229, 2019.

[169] J. C. Visier and R. A. Buswell, "Commissioning Tools for Improved Building Energy Performance," 2010.

[170] L. Dobiášová and J. Široky, "Versatile AHU fault detection – Design , field validation and practical application," *Energy Build.*, vol. 237, 2021.

[171] "ABB Cylon - Building and Home Automation Solutions | ABB." [Online]. Available: https://new.abb.com/low-voltage/products/building-automation/product-range/abb-cylon. [Accessed: 12-May-2023].

[172] "JACE 8000 Controller | Niagara | Tridium." [Online]. Available: https://www.tridium.com/us/en/Products/niagara/jace-8000. [Accessed: 12-May-2023].

[173] "OSIsoft - PI System | Operational Intelligence – Data Infrastructure." [Online]. Available: https://resources.osisoft.com/. [Accessed: 22-May-2023].

[174] "ASHRAE BACnet Committee Works with Other Organizations on New Standard | ashrae.org." [Online]. Available: https://www.ashrae.org/news/esociety/ashrae-bacnet-committee-works-with-other-organizations-on-new-standard. [Accessed: 22-Apr-2022].

[175] "Home – Project Haystack." [Online]. Available: https://project-haystack.org/. [Accessed: 24-Mar-2021].

[176] "Home - BrickSchema." [Online]. Available: https://brickschema.org/. [Accessed: 02-Mar-2022].

[177] "ASHRAE Standard 223P | Department of Energy." [Online]. Available: https://www.energy.gov/eere/buildings/ashrae-standard-223p. [Accessed: 12-May-2023].

[178] "About NREL | NREL." [Online]. Available: https://www.nrel.gov/about/. [Accessed: 12-May-2023].

[179] "Home – Lawrence Berkeley National Laboratory." [Online]. Available: https://www.lbl.gov/. [Accessed: 12-May-2023].

[180] "PNNL: Portland Office." [Online]. Available: https://www.pnnl.gov/portland/. [Accessed: 12-May-2023].

[181] "Resources." [Online]. Available: https://marketing.project-haystack.org/project-haystack-resources/white-papers. [Accessed: 09-Mar-2023].

[182] K. Bruton, P. Raftery, P. O'Donovan, N. Aughney, M. M. Keane, and D. T. J. O'Sullivan, "Development and alpha testing of a cloud based automated fault detection and diagnosis tool for Air Handling Units," *Autom. Constr.*, vol. 39, pp. 70–83, 2014.

[183] J. Schein, S. T. Bushby, N. S. Castro, and J. M. House, "A rule-based fault detection method for air handling units," vol. 38, pp. 1485–1492, 2006.

[184] "Welcome to Python.org." [Online]. Available: https://www.python.org/. [Accessed: 22-May-2023].

[185] "R: The R Project for Statistical Computing." [Online]. Available: https://www.r-project.org/. [Accessed: 22-May-2023].

[186] P. Raftery, A. Geronazzo, H. Cheng, and G. Paliaga, "Quantifying energy losses in hot water reheat systems," *Energy Build.*, vol. 179, pp. 183–199, 2018.

[187] H. Yang, S. Cho, C. S. Tae, and M. Zaheeruddin, "Sequential rule based algorithms for temperature sensor fault detection in air handling units," *Energy Convers. Manag.*, vol. 49, no. 8, pp. 2291–2306, 2008.

[188] E. A. Ashari, R. Nikoukhah, and S. L. Campbell, "Effects of feedback on active fault detection," *Automatica*, vol. 48, no. 5, pp. 866–872, 2012.

[189] T. Gao, B. Boguslawski, S. Marié, P. Bé, and S. Thebault, "Data mining and data-driven modelling for Air Handling Unit fault detection," no. June, pp. 2013–2017, 2019.

[190] H. H. Hosamo, P. R. Svennevig, K. Svidt, D. Han, and H. K. Nielsen, "A Digital Twin predictive maintenance framework of air handling units based on automatic fault detection and diagnostics," *Energy Build.*, vol. 261, p. 111988, 2022.

[191] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 327–346, 2003.

[192] Sustainable Energy Authority of Ireland, "Commercial / Industrial Fuels Comparison of Energy Costs," no. October 2022, pp. 54–55, 2022.

[193] J. G. Allen, P. MacNaughton, U. Satish, S. Santanam, J. Vallarino, and J. D. Spengler, "Associations of cognitive function scores with carbon dioxide, ventilation, and volatile organic compound exposures in office workers: A controlled exposure study of green and conventional office environments," *Environ. Health Perspect.*, vol. 124, no. 6, pp. 805–812, 2015.

[194] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nat. Rev. Phys.*, vol. 3, no. 6, pp. 422–440, 2021.

[195] "DENiM." [Online]. Available: https://denim-fof.eu/. [Accessed: 17-Apr-2023].

[196] G. Engels, R. Heckel, and S. Sauer, "UML – a universal modeling language?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1825, pp. 24–38, 2000.

[197] Martin Fowler, *UML Distilled Second Edition A Brief Guide To de Standart Object Modeling Language*. 2004.

[198] G. Rumbaugh, J., Jacobson, I., & Booch, *The Unified Modeling Language Reference Manual*. 1999.

[199] J. Rumbaugh, I. Jacobson, and G. Booch, *The Unified Modeling Language Reference Manual Second Edition*. .

[200] "What is UML | Unified Modeling Language." [Online]. Available: https://www.uml.org/what-is-uml.htm. [Accessed: 15-May-2023].

[201] "UML 2.5 Diagrams Overview." [Online]. Available: https://www.uml-diagrams.org/uml-25-diagrams.html. [Accessed: 15-May-2023].

[202] H. Koç, A. Mert, Y. Barjakly, and S. Peker, "UML Diagrams in Software Engineering Research : A Systematic Literature Review †," 2021.

[203] "RDF – Project Haystack." [Online]. Available: https://project-haystack.org/doc/docHaystack/Rdf. [Accessed: 24-May-2023].

[204] "Creating a Brick Model — Brick Ontology Documentation." [Online]. Available: https://docs.brickschema.org/lifecycle/creation.html. [Accessed: 24-May-2023].

[205] "What Is a Sandbox Environment? | CloudShare." [Online]. Available: https://www.cloudshare.com/virtual-it-labs-glossary/what-is-a-sandbox-environment/. [Accessed: 28-Apr-2023].

[206] "Containerization Explained | IBM." [Online]. Available: https://www.ibm.com/topics/containerization. [Accessed: 28-Apr-2023].

[207] "Heating & Cooling Degree Days – Free Worldwide Data Calculation." [Online]. Available: https://www.degreedays.net/. [Accessed: 07-Mar-2022].

[208] "update-11-feb-it-breach-cork-campus | Munster Technological University - MTU." [Online]. Available: https://www.mtu.ie/news/update-11-feb-it-breach-cork-campus/. [Accessed: 23-May-2023].