TALLINN UNIVERSITY OF TECHNOLOGY
SCHOOL OF ENGINEERING
Department of Electrical Power Engineering and Mechatronics

ROOM OCCUPANCY ESTIMATION USING ENVIRONMENTAL SENSORS

RUUMI HÕIVATUSE HINDAMINE KESKKONNAANDURITE ABIL

MASTER THESIS

Student: Prisca Adenike Adeoti
Student code: 213875MAHM
Supervisor: Kristina Vassiljeva, Associate Professor
Co- Supervisor: Daniil Valme, Early Stage Researcher

Tallinn, 2023

AUTHOR'S DECLARATION

Hereby I declare, that I have written this thesis independently.
No academic degree has been applied for based on this material. All works, major viewpoints and data of the other authors used in this thesis have been referenced.

"....." ..... 20……

Author: .....
/signature /

Thesis is in accordance with terms and requirements

"....." ..... 20…..

Supervisor: ….....
/signature/

Accepted for defence

"....."..... 20…. .

Chairman of theses defence commission: .....

/name and signature/

Non-exclusive Licence for Publication and Reproduction of Graduation Thesis[1]

I, Prisca Adenike Adeoti (name of the author) hereby

1. grant Tallinn University of Technology (TalTech) a non-exclusive license for my thesis

Room Occupancy Estimation Using Environmental Sensors,

(*title of the graduation thesis*)

supervised by Kristina Vassiljeva and Daniil Valme,

(*Supervisor's name*)

1.1 reproduced for the purposes of preservation and electronic publication, incl. to be entered in the digital collection of TalTech library until expiry of the term of copyright;

1.2 published via the web of TalTech, incl. to be entered in the digital collection of TalTech library until expiry of the term of copyright.

1.3 I am aware that the author also retains the rights specified in clause 1 of this license.

2. I confirm that granting the non-exclusive license does not infringe third persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

18/05/2023 (*date)*

[1] *The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.*

THESIS ABSTRACT

Author: Prisca Adenike Adeoti *Type of the work:* Master Thesis

Title: Room occupancy estimation using environmental sensors

Date: 18.05.2023 62 *pages (the number of thesis pages including appendices)*

University: Tallinn University of Technology

School: School of Engineering

Department: Department of Electrical Power Engineering and Mechatronics

Supervisor of the thesis: Associate Professor Kristina Vassiljeva

Co-supervisor: Daniil Valme

Abstract:

The achievement of energy efficiency and provision of inhabitants with comfortable indoor environments depend on the precise calculation of occupancy in buildings. To estimate occupancy, this study looks at how ventilation energy consumption, CO_2, and Temperature data can be used in conjunction with machine learning algorithms. To develop a viable dataset for the study, data had to be collected, cleaned, processed, and transformed. The machine learning techniques employed includes k-means clustering, DBSCAN, and Gaussian Mixture Models, and the effectiveness of each approach was assessed. This study's findings will contribute to the development of energy-efficient building control systems and improved indoor environment quality.

Keywords: Machine Learning, Clustering, CO_2 mass balance, Ventilation, Occupancy Estimation, Master Thesis

LÕPUTÖÖ LÜHIKOKKUVÕTE

Autor: Prisca Adenike Adeoti	*Lõputöö liik:* Magistritöö
Töö pealkiri: Ruumi hõivatuse hindamine keskkonnaandurite abil	
Kuupäev: 18.05.2023	18 *lk (lõputöö lehekülgede arv koos lisadega)*
Ülikool: Tallinna Tehnikaülikool *Teaduskond:* Inseneriteaduskond *Instituut:* Elektroenergeetika ja mehhatroonika instituut	
Töö juhendaja(d): Associate Professor Kristina Vassiljeva *Kaasjuhendaja*: Daniil Valme	
Sisu kirjeldus: Energiatõhususe saavutamine ja mugavate siseruumide tagamine sõltub hoonetes viibivate inimeste täpsest arvutamisest. Selleks, et hinnata hõivatust, uurib see uuring, kuidas ventilatsiooni energiatarvet, CO_2 ja temperatuuriandmeid saab masinõppe algoritmide abil kasutada. Usaldusväärse andmekogumi väljatöötamiseks koguti, puhastati, töödeldi ja muundati andmeid. Kasutatud masinõppetehnikad hõlmavad k-means klasterdamist, DBSCAN-i ja Gaussian Mixture mudeleid ning hinnati iga lähenemisviisi tõhusust. Selle uuringu tulemused aitavad kaasa energiasäästlike hoone juhtimissüsteemide ja parema siseruumide keskkonnakvaliteedi arendamisele.	
Märksõnad: Masinõpe, Klasterdamine, CO_2 massibalanss, Ventilatsioon, Hõivatuse hindamine, Magistritöö	

THESIS TASK

Thesis title:	**Room occupancy estimation using environmental sensors**
Thesis title in Estonian:	**Ruumi hõivatuse hindamine keskkonnaandurite abil**
Student:	**Prisca Adenike Adeoti, 213875MAHM**
Programme:	**Mechatronics**
Type of the work:	**Master Thesis**
Supervisor of the thesis:	**Assoc. Prof. Kristina Vassiljeva**
Co-supervisor of the thesis: (company, position and contact)	**Daniil Valme (Early Stage Researcher, Electrical Power Engineering and Mechatronics department)**
Validity period of the thesis task:	
Submission deadline of the thesis:	**May 18th 2023**

_____	_____	_____
Student (signature)	Supervisor (signature)	Head of programme (signature)
_____		
Co-supervisor (signature)		

1. Reasons for choosing the topic

This topic was chosen to support the mission of the European Green Deal which was approved in 2020. Additionally, the current energy crisis in Europe brought about by the Russion Federation aggression has also necessitated the creation of methods to reduce energy consumption. HVAC systems currently use up to 40% of the energy supplied to building as they currently work on schedule and not as required by the indoor human needs. Running building amenities according to the estimation of occupants in building will greatly reduce energy consumption. The reason for this work is to apply my interest in data analysis and machine learning to help reduce energy wastage and run building facilities optimally.

2. Thesis objective

The aim of this thesis is to develop an AI model with the best accuracy to predict occupancy level in building using already available environmental sensors in buildings thus reducing energy consumption while providing the required Indoor Air Quality (IAQ) for the comfort of building occupants.

3. List of sub-questions:

List 3-4 specific research goals that you intend to achieve or find and answer to.

- What environmental sensor data is required to estimate occupancy with highest accuracy?
- How does algorithm type affect accuracy of estimation model?
- How does ventilation flow rate and volume of room affect occupancy estimation?

4. Basic data:

The data for this work will be provided by the thesis supervisor

5. Research methods

This thesis work will use python for exploratory data analysis. The data will first be pre-processed to normalize the data and account for certain lags. The data condition will be a deciding factor on whether to roll data. Next, the clustering model will be developed using DBSCAN and K-Means. This model will be developed for different building types which include Mall, Kindergarten and Senior School building types. The results will be analyzed to understand how each sensor type affects the model accuracy.

Then finally, analysis will be carried out to estimate the possibility of using the same estimation model for different building types to access how much additional actions or fine-tuning is needed to apply it.

6. Graphical material

Algorithm blocks will be provided to visualise how the algorithms works. Tables and graphs of the various comparison and results will also be provided to further buttress the result of the work.

7. Thesis structure

This thesis report is organized into eight chapters.

Chapter 1 explains the thesis' goals in general terms and gives an introduction. There is an explanation of the problem description and the tasks required to accomplish those goals.

Chapter 2 presents an overview of relevant literature. It covers two main methods of occupancy estimation and explains why ML techniques are widely adopted and adopted in this work. It also explains the various data acquisition methods, their advantages, and disadvantages. It also highlights the ML method to be used in this work.

Chapter 3 describes the Data acquisition, preprocessing and analysis methods used.

Chapter 4 discusses the development and analysis of the model using various metrics for example accuracy, RMSE, specificity, etc.

Chapter 5 provides ideas on future research works and recommendations based on the findings of the study.

Chapters 6 provides ideas on future research works and recommendations based on the findings of the study in Estonian

Chapter 7 lists the references used in this study.

Chapter 8 lists the appendices

8. References

Literatures from researchgate and library will be used

- Dong, B., Andrews, B., Lam, K. P., Höynck, M., Zhang, R., Chiou, Y.-S., & Benitez, D. (2010). An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network. Energy and Buildings, 42(7), 1038–1046. https://doi.org/10.1016/j.enbuild.2010.01.016
- Simma, K. C. J., Mammoli, A., & Bogus, S. M. (2019). Real-time occupancy estimation using WiFi network to optimize HVAC operation. Procedia Computer Science, 155, 495–502. https://doi.org/10.1016/j.procs.2019.08.069.
- Wang, W., Chen, J., & Hong, T. (2018). Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings. Automation in Construction, 94, 233–243. https://doi.org/10.1016/j.autcon.2018.07.007
- Ding, Y., Han, S., Tian, Z., Yao, J., Chen, W., & Zhang, Q. (2022). Review on occupancy detection and prediction in building simulation. In Building Simulation (Vol. 15, Issue 3, pp. 333–356). Tsinghua University. https://doi.org/10.1007/s12273-021-0813-8
- Zuraimi, M. S., Pantazaras, A., Chaturvedi, K. A., Yang, J. J., Tham, K. W., & Lee, S. E. (2017). Predicting occupancy counts using physical and statistical Co2-based modeling methodologies. Building and Environment, 123, 517–528. https://doi.org/10.1016/j.buildenv.2017.07.027
- Shengwei Wang and Xinqiao Jin, "CO2-based occupancy detection for on-line outdoor air flow control," Indoor and Built Environment, vol. 7, no. 3, pp. 165–181, May 1998, doi: 10.1177/1420326X9800700305

9. Thesis consultants

N/A

10. Work stages and schedule

Analysing relevant literatures (November 2022)
Gathering required data and Data pre-processing (December 2022)
Model creation using AI algorithms (January 2023)
Comparison of model accuracy with respect to building type, analysing the effect of different data on model accuracy and describing results of the study (February 2023)
Compiling the theoretical part of the thesis work and sending to supervisor for first reading (March 2023)
Effecting corrections and sending to supervisor for second reading and completing the final version of the thesis (April 2023)
Buffer period and submission (May 2023)

TABLE OF CONTENTS

PREFACE

I would like to express my sincere gratitude to God Almighty for giving me the strength and privilege to complete this thesis work successfully, especially during the most challenging time of my life.

I would also like to extend my heartfelt appreciation to my supervisor, Kristina Vassiljeva, for her guidance, support, and invaluable contributions towards the completion of this research work. Without her constant willingness to create time and direct me to relevant materials, this thesis would not have been possible. I also want to acknowledge my co-supervisor, Daniil Valme.

I am deeply grateful to my family, friends, and well-wishers for their unwavering support throughout my academic journey. To my siblings, Terah, Kristin, Edna, Emmanuella, Daniel, and Quincy, thank you for your prayers and encouragement during this period. Lastly, I want to give a special thanks to Verrif, for providing me with the necessary tools and resources to learn Data Analysis and Machine Learning. A very big thank you to Raymond Aigbe Ogbebor, you were a very big inspiration to me on my journey and I will never forget it.

Once again, thank you all for your support, encouragement, and contributions towards the successful completion of this thesis.

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AER	Air Exchange Rates
AHU	Air Handling Unit
AI	Artificial Intelligence
ANN	Artificial Neural Network
BIM	Building Information Modeling
BMI	Body Mass Index
BMS	Building Management System
BN-ANN	Building Network Artificial Neural Network
BP	Back Propagation
CO_2	Carbon dioxide
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DCV	Demand Controlled Ventilation
EU	European Union
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Model
HVAC	Heating, Ventilation and Cooling
IAQ	Indoor Air Quality
IoT	Internet of Things
ITCNN	Intelligent Thermal Comfort Neural Network
K-NN	K-Nearest Neighbor
LT	Lecture Theatre
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
NDIR	Non-Dispersive InfraRed
PEM	Prediction Error Minimization
PPM	Parts Per Million
PIR	Passive Infrared
PTZ	Pan-Tilt-Zoom
RBF	radial basis function
REHVA	Federation of European Heating, Ventilation and Air Conditioning Associations
ReLU	Rectified Linear activation Unit
RF	Random Forest

RH	Relative Humidity
RL	Reinforced Learning
RMSE	Root Mean Squared Error
SNA	Social Network Analysis
SVM	Support Vector Machine
TCV	Thermal Comfort Votes
TSV	Thermal Sensation Votes
WCSS	Within-Cluster Sum of Square
Wi-Fi	Wireless Fidelity
WSN	Wireless Sensor Network

1. INTRODUCTION

Buildings play a significant role in human daily lives because we spend the majority of our time inside of them [1]. They include residential and commercial structures, workplaces, libraries, shopping centers, hospitals, schools, and other public buildings. It is necessary to use electricity-powered utilities such as lighting, heating, ventilation, etc. while occupying these buildings. According to data from the U.S. Department of Energy, HVAC account for 35% to 45% of all maintenance costs in a building [2]. In the EU, HVAC accounts for 38% of energy use in buildings according to the European commission website. To keep these structures habitable, HVAC systems are a necessity in public buildings. Even on weekends and holidays when the buildings are empty and do not require as much ventilation as when they are inhabited to capacity, these HVAC systems are operated at full capacity.

The advancement of technology has facilitated the collection of a large amount of data in various fields, including the field of building automation. With the increasing focus on energy conservation and building efficiency, many researchers have keyed into the application of machine learning techniques in building automation. One of the key challenges in building automation is to estimate the occupancy of a room accurately.

Traditionally, occupancy estimation has been performed using motion sensors or manual counting. However, these methods are not always reliable and carry along many inaccuracies. Furthermore, the use of cameras for occupancy estimation raises privacy concerns. As a result, there has been a growing interest in the use of environmental sensors for occupancy estimation.

Environmental sensors can measure various parameters such as temperature, humidity, and CO_2 levels, which are affected by the presence of occupants in a room. Machine learning algorithms can then be used to analyze these sensor readings and estimate the occupancy of the room. This approach has several advantages over traditional methods, including its non-intrusive nature, low cost, and ease of installation.

1.1 Problem statement

DCV aims to reduce the amount of heating and cooling required by buildings by altering the ventilation rates according to occupancy levels. These systems generally depend on the concentration of selected markers and results in less energy consumption without affecting the acceptable IAQ [3]. To this end, it is essential to accurately determine the occupancy levels in a space to properly implement DCV. Estimating occupancy using counting gates is expensive to deploy at room level inside a whole building [4], and hence the goal of this thesis is to develop an AI model system that can operate on already installed equipment in buildings to reduce energy consumption by estimating or predicting the occupancy level at a specific moment and operating energy systems based on that knowledge which can then be used in BMS and BIM services for building.

This study attempts to investigate solutions to assist the present energy crisis, especially in the EU due to the conflict, which has made it imperative to explore strategies to reduce energy consumption in buildings without incurring additional costs due to the installation of new infrastructure. To achieve this goal, accurate and reliable occupancy estimation methods are required for building automation. However, traditional methods such as motion detection or door sensors have limitations in terms of accuracy and practicality. This study proposes a new method for occupancy estimation using environmental sensors, specifically temperature, and CO_2 sensors. The objective is to develop machine learning algorithms that can analyze sensor readings and accurately predict room occupancy. The study aims to evaluate the performance of the proposed method using real-world data collected from several buildings. The outcome of this study will contribute to the development of more accurate and reliable occupancy estimation methods for building automation, leading to increased energy efficiency, reduced environmental impact, and improved comfort for building occupants.

1.3 Description of task

The list of tasks that will be done during this study includes:

I. Overview of supporting literature on data that correlates with occupancy counts and use of environmental sensors for estimating occupancy count.
II. Data pre-processing to clean the data to increase overall productivity and allow for the highest quality information of models
III. Model creation using data clustering
IV. Analysis of models created
V. Validation of models using building types.

1.4 Out of scope of this work

While the proposed topic was motivated by an interest in data collection, data analysis, data science, and energy price reduction, this work does not focus on the practical implementation of the proposed algorithm. The hardware application of the proposed algorithm is not within the scope of this thesis.

This work does not aim to develop or implement the sensors, or the sensor network required for occupancy estimation, or to for example, implement the model to an HVAC system in the building. Instead, the focus is on analyzing and building models that can be provided to construction companies to be integrated into building information modeling (BIM) and building management systems (BMS) for large buildings. The goal of this study is not to provide hardware solutions, but rather to investigate how data affects the model and what combination of data is best suited for a particular circumstance. The study also aims to identify the achievable results given a set of data.

1.5 Thesis structure

This thesis report is organized into eight chapters.

Chapter 1 explains the thesis' goals in general terms and gives an introduction. There is an explanation of the problem description and the tasks required to accomplish those goals.

Chapter 2 presents an overview of relevant literature. It covers two main methods of occupancy estimation and explains why ML techniques are widely adopted and adopted in this work. It also explains the various data acquisition methods, their advantages, and disadvantages. It also highlights the ML method to be used in this work.

Chapter 3 describes the Data acquisition, preprocessing and analysis methods used.

Chapter 4 discusses the development of the ML models and analysis of the model using CO_2 mass balance equation.

Chapter 5 provides ideas on future research works and recommendations based on the findings of the study.

Chapters 6 provides a summary of the thesis.

Chapter 7 lists the references used in this study.

2. THEORETICAL BASIS

2.1 Background of work and literature review

This section provides an overview of relevant literature on the prediction of occupancy for the reduction of energy consumption in buildings. The terms "occupancy" and "occupant behavior" refer to the presence of people inside structures and their active engagement with various building systems, including lighting, heating, cooling, ventilation, window coverings, plugs, etc. [5]. Occupancy information is vital in the design, operation, and energy efficiency of buildings. Occupancy information may be used in pattern identification, behavior prediction, sensing and tracking, occupancy detection, and quantitative prediction. First, the mass balance equation method for occupancy estimation will be discussed. The shortfall of this method of occupancy estimation will be mentioned. Second, the use of ML techniques as an alternative will be considered. Finally, different data acquisition methods for model creation, their advantages and disadvantages and some metrics for model accuracy estimation will be discussed.

2.2 Traditional techniques for occupancy estimation

The traditional techniques for estimating occupancy in smart buildings typically rely on motion sensors. These sensors detect movement within a certain range and assume that any movement corresponds to the presence of a person. However, there are limitations to this approach. For example, if a person is stationary or moves slowly, the sensors may not detect their presence, leading to an underestimation of occupancy. On the other hand, if there are objects or pets that move within the range of the sensors, they may be mistakenly identified as people, leading to an overestimation of occupancy.

The use of opportunistic sensor data, specifically motion sensors installed by security companies, to infer the number of residents in a house and their identities [6]. And as explained previously, the model showed some inaccuracies The authors then explored the possibility of using machine learning algorithms to discriminate movement trajectories of different occupants to identify the current occupants based on anonymous motion events.

2.3 Environmental sensors data as a means of occupancy estimation

To implement DCV, it is necessary to accurately determine the occupancy of a space, and this can be achieved by processing data from a plethora of sources including cameras, gas sensors, Wi-Fi, CO_2 sensors, Temperature sensors, etc. Cost factors and privacy concerns have led to the widespread usage of CO_2 sensors. The foundation for employing CO_2 for occupancy estimation is based on well-quantified principles of human physiology. All humans exhale CO_2 at a consistent rate based on occupant age and activity level while engaged in similar levels of activity [7].

According to earlier research, some environmental sensors exhibit a significant correlation between their values and building occupancy rates. The viability and possibility of using specific sensor data and analysis techniques for occupancy prediction were examined in [8]. The authors applied environmental sensors to office buildings for occupancy estimation at the Robert L. Preger Intelligent Workplace (IW) at Carnegie Mellon University. They created a comprehensive, all-encompassing environmental sensing testbed that included distributed sensors for a range of environmental parameters like CO_2, carbon monoxide (CO), total volatile organic compounds (TVOC), small particulates (PM2,5), acoustics, illumination, motion, temperature, and relative humidity. According to their findings, there was a strong correlation between occupant count and CO_2 and acoustic parameters. However, difficulties can occur when using acoustics because of the impact of sound from neighboring offices. Environmental sensors, including luminance, temperature, relative humidity (RH), motions, CO_2 concentration, power consumption, door and window positions, and acoustic pressure from microphone sensors were used in [4] to collect data to estimate occupancy.

It was discovered that door opening contacts, motion (PIR) sensors, power consumption sensors, CO_2 sensors, a microphone, and power consumption sensors had the strongest correlations with occupancy. Estimation of occupancy form environmental data could be done using the mass balance method, machine learning method, etc. This chapter discusses mass balance and Machine learning methods.

2.4 Indoor CO_2 concentration standards and guidelines

Indoor CO_2 levels can pose a threat to human health well before they reach 5.000 ppm. As a result, various standards and guidelines have been put in place to ensure that indoor air quality remains within acceptable limits. REHVA has established guidelines related to indoor CO_2 levels in various types of buildings, including schools, offices, and residential buildings [9]. The recommended CO_2 levels in these guidelines vary depending on the specific building type and occupancy. For example, in office buildings, REHVA recommends that CO_2 levels should not exceed 1.000 ppm above outdoor levels, and that ventilation rates should be sufficient to maintain CO_2 levels below this threshold. In classrooms and other educational buildings, REHVA recommends a maximum CO_2 level of 1,500 ppm, while in residential buildings, the recommended maximum CO_2 level is 1,200 ppm.

2.5 Quantitative occupancy prediction

Occupancy prediction can be classified under detection or estimations. Occupancy detection involves detecting the presence or absence in a space while occupancy prediction is about estimating the number of occupants in a space. This research focuses on occupancy estimation to reduce energy consumption. In the work [5], a correlation between CO_2 concentration and occupancy was investigated using a synthetic variable defined as the volume available per person. They opined that data on CO_2 levels has throughout time shown relationships with several factors, including occupancy profile, since CO_2 levels rise proportionally to the level of activity because of metabolic activities.

To quantify CO_2 concentrations in enclosed environments, defining a few numerical details is crucial. In businesses and learning environments like schools and colleges, a level of 600-1.000 ppm is deemed ideal.

Physical and statistical models were employed in [10] to predict the occupancy counts in a high-volume lecture theatre. They collected 6189 datasets spread across a four-month period and used a PTZ camera to obtain ground truth by recording images at 5-minute intervals. They used 3 CO_2 sensors spread throughout the room to collect data at 5-minute intervals and took an average of it to account for brief increases in CO_2 brought on by people breathing directly on the sensors. The physical model utilized the fully mixed dynamic mass balance model is given in eq. (2.1).

$$V \cdot \frac{dCO_{2,in}}{dt} = \lambda_v \cdot CO_{2,out} + R - \lambda_v \cdot CO_{2,in} \quad (2.1)$$

Where $CO_{2,in}$ - indoor CO_2 concentration in the LT, ppm/10^6,

$CO_{2,out}$ - outdoor CO_2 concentration, ppm/10^6,

λ_v - AER of the LT,h^{-1},

R - rate of CO_2 being generated within the LT, $Lmin^{-1}$,

V - volume of the LT, m^3.

One significant flaw in this approach is that both models experience residual CO_2 concentration after people have left the room. This can be mitigated by coupling a CO_2 sensor with, for example, a PIR sensor.

2.6 CO_2 mass balance method for occupancy estimation

This estimate is based on a model of the dynamics of CO_2 gas inside a well-mixed environment, with the assumption that the air mass in the environment is constant and the concentration distribution in the environment is spatially uniform. The work [7] gives the following mass balance equation to describe the flow of gas in an enclosed space described by Figure 2.1:

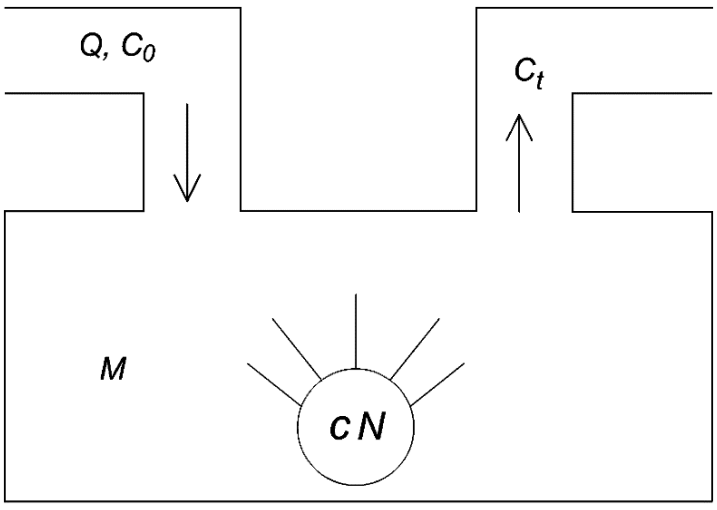


Figure 2.1 Hypothetical indoor space [7]

$$V\dot{C}t = -Q(Ct\text{-}C0) + cN \quad (2.1)$$

The unknowns that will be estimated are to be the number of occupants N and the airflow rate Q. For this work, mass, M will be replaced with volume, V. The eq (2.2) results from integrating this equation and presuming there is not already a gas source in the space:

$$C_t = C_0 + \left(\frac{cN}{Q}\right)\left(1 - e^{-\frac{tQ}{V}}\right) \quad (2.2)$$

Where C_t - CO_2 concentration in the space at a particular time t, mgm^{-3},

C_0 - CO_2 concentration in the space at a particular time t = 0, mgm^{-3},

c - CO_2 generation rate per person in the space, mgh^{-1},

N - number of people in the space at time t,

Q - volumetric flow rate of the ventilation unit, m^3/h

t - time under consideration, h)

V - volume of the building or space being considered, m^3

Note:

$$Mgm^{-3} = 0.0409 * ppm * CO_2 \text{ molar weight } (28.01) \quad (2.3)$$

Theoretical CO_2 mass balance equation-based techniques given above are effective at forecasting occupant counts in the order of tens, but they require the user to submit numerous details about the observed rooms, such as volume and air flow rate, or the results of air flow meters [11]. The main flaw of this method is that not all parameters can always be measured or obtained in buildings. Most buildings do, however, include environmental sensors that can be used in conjunction with other techniques, such Machine Learning (ML) techniques, which are covered below, to predict occupancy.

2.7 Machine learning methods for occupancy estimation

Computer systems use algorithms and statistical models to carry out certain tasks without being explicitly programmed. This process is known as Machine Learning (ML) and is also a subset of Artificial Intelligence (AI). A machine learning algorithm is a computer procedure that uses input data to complete an intended goal without being explicitly programmed to do so. These algorithms are, in a way, "soft programmed" in that they automatically change or adjust their design because of repetition to get better and better at carrying out the desired task. Training is the process of adaptation, when samples of the input data are given along with the intended results. The algorithm is then set up in the best possible way so that it can provide the desired result when given the training data as well as generalize to create the desired result from fresh, previously unexplored data. The "learning" component of machine learning is this training. The training does not have to be restricted to a first adaptation over a set period. A smart algorithm can engage in "lifelong" learning as it analyses fresh data and learns from its errors, just like humans can [12].

Many of the applications that we use every day have learning algorithms [13]. As shown in Figure 2.2, ML algorithms can be classified into various classes. The authors of [14] proposes a new approach to estimating occupancy in smart buildings using CO_2 measurements. The authors argue that traditional occupancy estimation techniques, such as motion sensors, are limited and can be inaccurate. In contrast, the authors propose a machine learning-based approach that uses CO_2 levels as a proxy for occupancy estimation. The idea behind this approach is that when people occupy a room, they exhale CO_2, which increases the concentration of CO_2 in the air.

By measuring the concentration of CO_2, the occupancy of the room can be estimated. The authors collected CO_2 and occupancy data in real buildings and used it to train machine learning models. The models were then tested on data collected from other buildings to evaluate their accuracy. The advantage of this approach is that it is not affected by motionless or slow-moving occupants, and it can distinguish between people and other objects that do not emit CO_2.

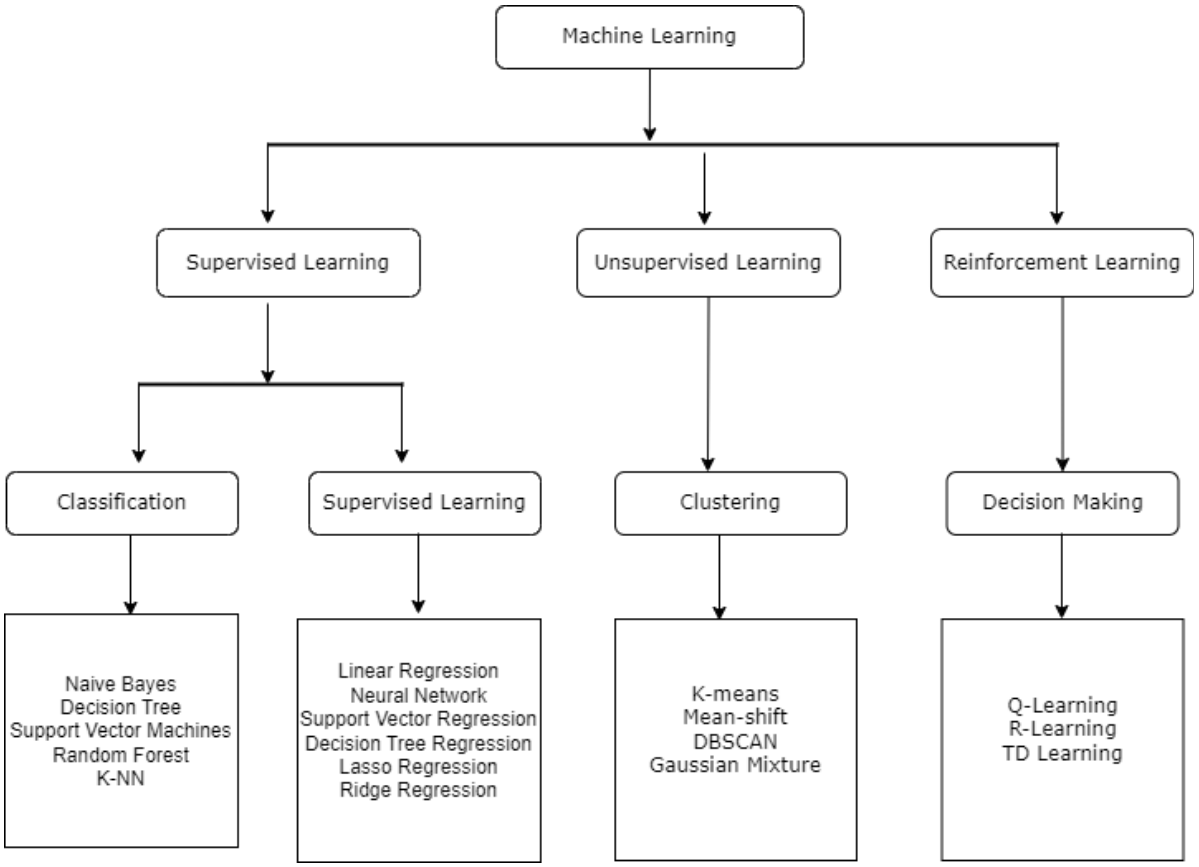


Figure 2.2 ML classification

2.7.1 Supervised Learning

Supervised Machine Learning is the pursuit of algorithms that draw broad hypotheses from instances supplied externally, and then forecast instances in the future [15]. In this type of ML technique, the machine is trained using "labeled" training data, and then they make predictions about the outcome using that data. "Labeled data" refers to input data that has already been given the correct output.

In supervised learning, the training data provided to the computers acts as the supervisor, teaching them how to accurately predict the output in a manner similar to how a student learns with a teacher. The supervised learning method involves giving the ML model the right input and output data. A supervised learning method looks for a mapping function to connect the input variable (x) with the output variable (y). Following the completion of the training phase, the model is evaluated using test data (a subset of the training set), and it then makes output predictions.

2.7.2 Unsupervised Learning

In unsupervised learning, algorithms are used to find patterns in data sets which have data points that are neither categorized nor assigned labels. Thus, without any outside assistance, the algorithms are free to classify, label, and organize the data points inside the data sets. The model does not require supervision from the users. In other words, the learning algorithm is not provided any labels, therefore it is left to its own devices to identify structure in the input collection. Unsupervised learning is when an AI system groups unsorted data based on similarities and differences even if no categories are given. Finding hidden and intriguing patterns in unlabeled data is its primary objective [16]. The four classes of unsupervised learning are: Clustering, Association, Anomaly detection, and Auto-encoders. This work focuses on clustering. A typical Machine Learning process can be diagrammatically represented as shown in Figure 2.3. As it relates to this work, ML algorithms have a series of applications in building management system field as can be seen from Figure 2.4.

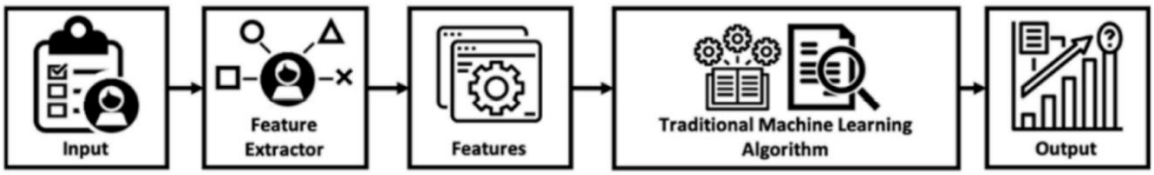


Figure 2.3 Flowchart of a typical ML process [17]

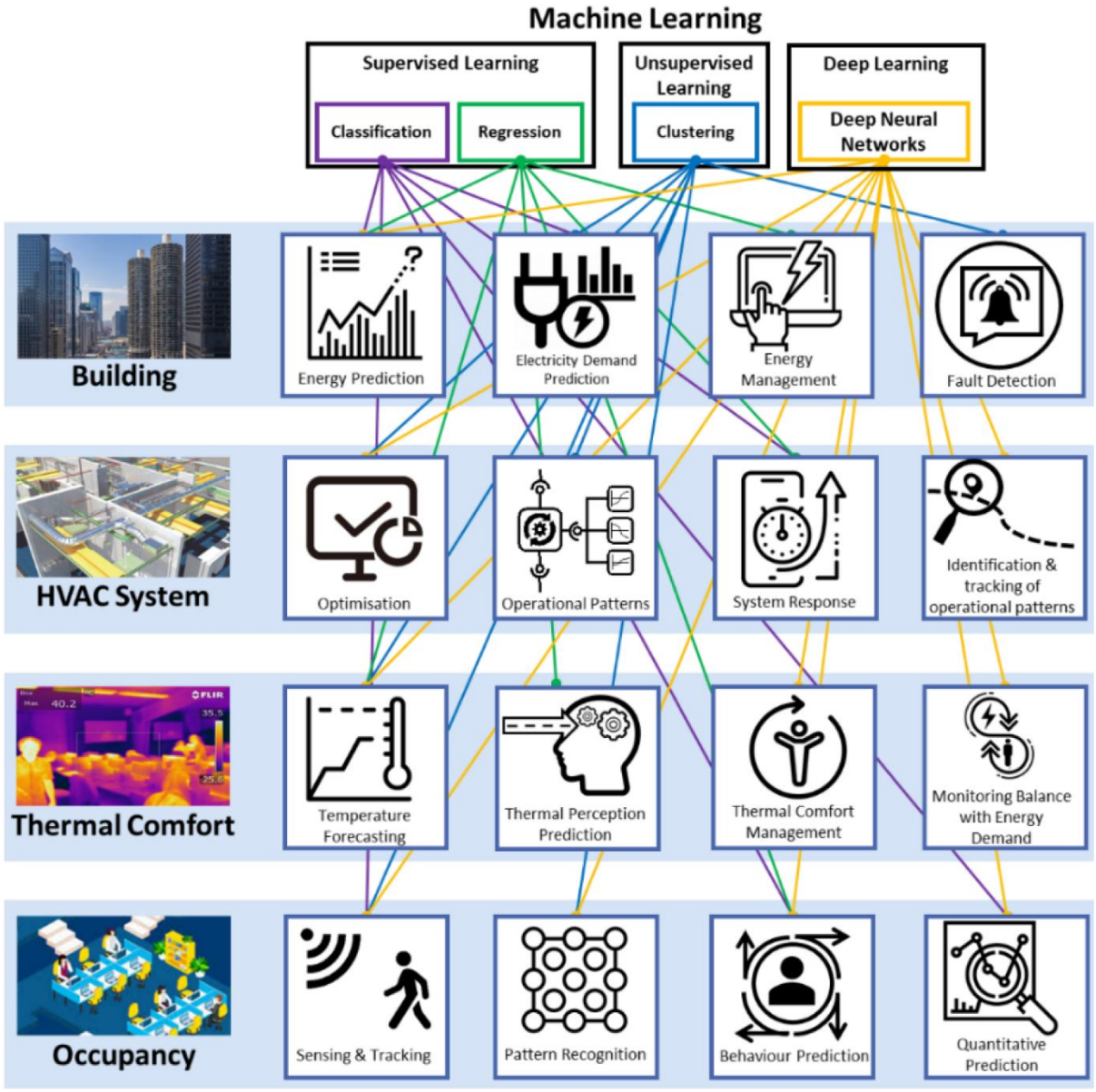


Figure 2.4 ML application as discussed in this chapter [17]

2.8 Data Acquisition methods for occupancy prediction using Machine Learning models

The environment has the power to improve or worsen people's quality of life [17]. The building industry has benefited from the continued advancements of AI and ML for smart buildings focusing on efficiency, thermal comfort, health, and productivity. These technologies are used in "smart" buildings to automate activities like lighting, HVAC, and security [18]. Due to their capacity to recognize patterns in data, machine learning algorithms have become effective tools for occupancy prediction. Wi-Fi, video, and environmental sensor data, as well as other forms of data, can all be utilized to estimate occupancy using machine learning algorithms. By utilizing the fact that occupancy patterns are frequently connected to changes in certain environmental parameters, it is possible, for instance, to use environmental sensor data such as CO_2 and temperature to anticipate occupancy. Video data can be used to forecast occupancy by recognizing the presence of individuals in the visual area, while Wi-Fi data can be used to predict occupancy by recording the movement patterns of people.

This section aims to provide an overview of the various data acquisition techniques for occupancy prediction using machine learning applications. We will discuss the advantages and limitations of different types of data sources and Machine Learning techniques for occupancy prediction. Energy prediction, Electricity demand prediction, Energy management and Fault detection are a few of the ways in which AI has been adapted in the built environment sector.

2.8.1 Using WI-FI data

The capacity to collect fine-grained human activities and the ubiquity of Wi-Fi networks have made occupancy estimation using Wi-Fi data a viable method in recent years. There are several theories on which this method works.

Firstly, because Wi-Fi signals are weakened by objects, such as people, Wi-Fi data can be used to estimate occupancy. This implies that alterations in the Wi-Fi access point signal intensity can be utilized to deduce the presence of people nearby.

Wi-Fi signals can also be used to determine users' locations, which can provide more details to the occupancy estimation process. Another theory is the fact that the number of people connected to the Wi-Fi network of a building can be counted which can give an estimate of the number of people in a building at a particular instance. This information can be used to train a ML model to predict occupancy at for energy reduction models. To train a classifier, for instance, to discriminate between signal patterns pertaining to empty and occupied locations, supervised learning techniques can be utilized. On the other hand, unsupervised learning methods can be used to cluster signal patterns and identify occupancy patterns. The authors of [19] applied Wi-Fi probe technology set to 30 seconds interval to collect Wi-Fi data and fused it with environmental sensor data (indoor air temperature, RH, and CO_2 concentration) to predict occupancy level using BP based ANN (with 3 hidden layers), SVM (with $\varepsilon = 0.2$) and K-NN (with $k= 15$) ML algorithms. The fusion was done using time label and ground truth was acquired using two overhead cameras installed to record the entrance and exit events of occupants and the occupant number was then manually counted after the recording. The experiment was conducted in a graduate student office inside a building housing academic facilities at City University of Hong Kong. Three data groups were used in the study to create three occupancy models, one using only environmental parameters (T, RH, and CO_2), one using only Wi-Fi data, and one utilizing both datasets. The MAE, MAPE, and RMSE indices were used to evaluate the occupancy models. They discovered that the ANN-based occupancy model outperforms the kNN and SVM models at predicting occupancy, not just when using environmental factors alone but also when using environmental parameters and Wi-Fi data together. The SVM model performed most effectively with the Wi-Fi data.

In the work [20], due to added infrastructure drawbacks, a simplistic framework was developed. This was based on the use of commodity Wi-Fi to estimate real-time occupancy data that can result in a large energy savings in HVAC operation while eliminating privacy concerns. The experiment was carried out in a large lecture hall in the Mechanical Engineering department at the University of New Mexico for six weeks. They used data from a sensor installed on the room door frame that counts the number of people in the room (ground truth) and router data that shows the number of people connected to the campus network inside the room. The main drawback of this strategy is the employment of camera and Wi-Fi technology, which raises privacy issues, as well as the possibility that a single person could have multiple wearable devices connected to the Wi-Fi router at a particular time giving false information. This method also falls short for this study as students in the school are not allowed to use their phones in classes during sessions.

2.8.2 Using measured vital signs

Previously, many researchers based their theory for thermal comfort prediction on Fanger's PMV model [21]. A ML algorithm was utilized in [22] to forecast residents' thermal comfort votes (TCV) and thermal sensation votes (TSV) in fourteen Chinese cities. Sensors were used to measure relative humidity, air speed, inside and outdoor air temperature, and globe temperature. Thermal comfort (TCV) and thermal sensation (TSV), metabolism, gender, BMI, age, climate type, level of clothing, and adaptive control measures such as opening windows or doors, using electric heaters, etc., was obtained using paper-based surveys. 10.000 dataset was obtained of which 5.512 were obtained in naturally ventilated spaces which were all normalized to ensure data accuracy. SVM and BP ANN algorithm was used to develop a model to predict TSV and TCV. The experiment revealed that for TCV prediction, the combination of indoor and outdoor environmental parameters, personal parameters, climatic types, and adaptive control measures revealed best result; while the combination of indoor and outdoor environmental parameters, personal parameters, climatic types, and adaptive control measures revealed best result for TSV prediction. However, Personal characteristics including age, gender, and BMI have little bearing on predicting thermal comfort of spaces. ML and IoT-based method as against Fanger's PMV method was used to access thermal comfort management as proposed in [23]. The authors used a WSN with various environmental sensors embedded in the wireless node connected to a common gateway used to transmit data to backend servers via the internet. A mobile app was also developed to track occupant vital signs via wearable devices in real time. The backend system was used to manage and analyze data. Thermal comfort modelling was carried out using 2 methods which are the white-box ML approaches; Native Bayes, K-NN, and decision tree and the black-box ML approaches; occupant dynamics, complex relationships, interaction among parameters, and confounding factors using SVM, RF and ITCNN. Using the information gathered from 20 students and 10 staff volunteers over a three-week period, they first created a simulation environment based on statistical theory. The research concluded that ITCNN outperforms the PMV model and the other six traditional machine learning techniques in terms of modeling performance. The drawback to this method is also privacy as stipulated by GDPR. Collection of data such as age can pose serious privacy non-compliance.

2.8.3 Using Video camera and motion sensor

Another means of data acquisition for model creation is using motion sensors and video cameras. In the work [24], an occupancy prediction model based on real-time occupancy data obtained from a wireless sensor network, Smart Camera Occupancy Position Estimation System (SCOPES) using the Markov Chain model was proposed. They concluded that to save energy, it is important to have real-time data on occupancy which they acquired from video data. Their study resulted in an average of 42% annual energy savings. The authors of [25] predicted occupancy using indoor data derived from sensors for various environmental parameters such as temperature, humidity, CO_2, illuminance, and motion. In addition to these, a web-based camera was also included in the sensing networks to track the exact occupancy in the test space. While the method described was found to be effective, there are also concerns about privacy. The building being studied mainly houses children, and collecting ground truth data through video surveillance would be intrusive and impractical.

2.8.4 Using combination sensors

Energy management is crucial for smart buildings and cities since it lowers power usage and results in improved energy and cost savings. Different algorithms have been developed over time to better manage energy consumption for built environments. The work [26] uses a dataset obtained from a nearby airport which contains 35 different variables of weather information (temperature, humidity, pressure, wind speed, visibility, and dew point), appliance and light energy consumption and temporal data. The dataset was recorded for 137 days at 10 minutes intervals. First, the dataset was normalized before being used in various configurations to train the model. The AI technique employed was a multi-layer feed-forward (MLP) neural network with a ReLU activation function with the output layer being the power usage. Various configurations were tested, and the optimum model had four hidden layers with 512 neurons per layer. This model had an RMSE of 66.295%, R^2 value of 56.7%, MAE of 29.556 and MAPE of 27.961.

In the paper [27], Data was collected from 2015 to 2017 in the Southeast University in China. The data was used to develop a model that integrates a SNA with a BN-ANN and ANN technique to predict multi-building energy use. ANN of 4 layers having a random number of hidden layers and a sigmoid activation function was used to predict energy use. It was concluded that the BN-ANN model offered better prediction accuracy compared to ANN. In the work [28], a data-driven model of a HVAC thermal comfort-based temperature set point control using k-nearest neighbor (KNN) with K=100, random forest (RF) with tree depth of 3, and support vector machine (SVM) with a degree of correctness C = 1, was developed. Eight features were extracted from the dataset which included the average of three heights' air temperature (°C), outdoor average min/max relative humidity on the day of survey (%), Average metabolic rate of subject (met), Relative humidity (%), outdoor average of min/max air temperature on day of survey (°C), Clothing plus upholstery insulation (clo), average of three heights' air speed (m/s), and Average of three heights' mean radiant temperature (°C). They used KNN, SVM and RF to classify the data and adopted a Q-learning controller algorithm as the temperature set point controller. The Q-learning algorithm is a reinforcement learning algorithm. Algorithms for reinforcement learning develop their own autonomous responses to their surroundings. Its agent attempts to maximize a numerical reward (gained from a correct output) signal through trial and error as it learns how to relate situations to actions. In this manner, the algorithm develops over time [29]. When there are several distinct types of sensors accessible, this strategy can be used. Unfortunately, this isn't often the case in most buildings, particularly in offices and schools. This makes it a less than ideal method of gathering data for ML model development.

Various data acquisition means have been employed in research literatures. Table 2.1 lists a few data acquisition means as they relate to building automation, their advantages, and drawbacks. This work will utilize data from CO_2, and temperature sensors as these environmental sensors are readily available in buildings by standard and are also non-intrusive, hence does not pose privacy threat to the students who are children.

Table 2.1 Table of data acquisition means showing their pros and cons [29], [30]

Data acquisition tool	Advantages	Drawbacks
PIR	Easy to deploy. Low cost. Non-intrusive. Easy detection.	Unable to detect static state multi-occupant situation. Limited detection range.
Ultrasonic	Detects minor motion. Does not require an unobstructed line of slight.	High levels of vibration or airflow complicates their application.
RFID	Easy to deploy. Real-time response.	Subject to indoor electromagnetic condition. Require users to carry a card/tag.
Camera	High accuracy. Missing data entries can be handled successfully.	High cost. Privacy problems. Complex model or algorithm required for processing data.
Temperature, RH, CO_2 sensor (so called environmental sensors)	Easy to deploy. Low cost. Commercially available. Non-intrusive.	Stability of the sensor is dependent on regular calibration. Delay problems.
GPS, Wi-Fi, Bluetooth	Efficient and convenient.	The join point does not match the number of occupants. It may require other sensors for effective occupancy detection. Privacy concerns. Requires a device to be carried by the occupant.
Combination sensor	High accuracy.	High cost.

2.9 Clustering as an unsupervised learning algorithm

Clustering, which is an exploratory technique, is used to locate dense groupings of data that are more comparable to one another. Quantifying the degree of similarity or dissimilarity between observations is necessary for this process. The type of similarity metric employed greatly affects the analysis's findings. It has numerous uses in pattern identification, picture analysis, consumer analytics, market segmentation, social network analysis, and other fields. From airplanes to healthcare and beyond, a wide range of businesses use clustering.

One of the main benefits of clustering over supervised learning is that it is unsupervised learning, meaning that we do not need labeled data for clustering algorithms. This work utilizes this learning method due to the absence of output labels. This was done to protect the privacy of the residents as our data does not include ground truth or video camera and occupancy information. The algorithms utilized in this work for model creation are K-Means, GMM and DBSCAN clustering.

2.9.1 K-means clustering

The K-means algorithm often uses the standard of square error and identifies K clusters in accordance with a particular standard [31]. The algorithm's foundation is the internal distance minimization (the sum of the distances of the patterns assigned to a cluster to the centroid of that cluster). The fundamental principle of K-means clustering is to move each point to its new nearest center if the initial clustering is not optimal, update the clustering centers by calculating the mean of the member points, and repeat the moving-and-updating process up until the convergence criteria (such as a predetermined number of iterations, a difference in the value of the distortion function) are met [32]. K-means algorithm uses a first set of centroids that are chosen at random to serve as the starting points for each cluster as it processes the learning data. Iterative computations are then performed to optimize the placements of the centroids. It stops developing and enhancing clusters when either the centroids have stabilized; their values have not changed, or iterations have reached the predetermined number. The K-means clustering technique is graphically represented in Figure 2.5 as a flow chart.

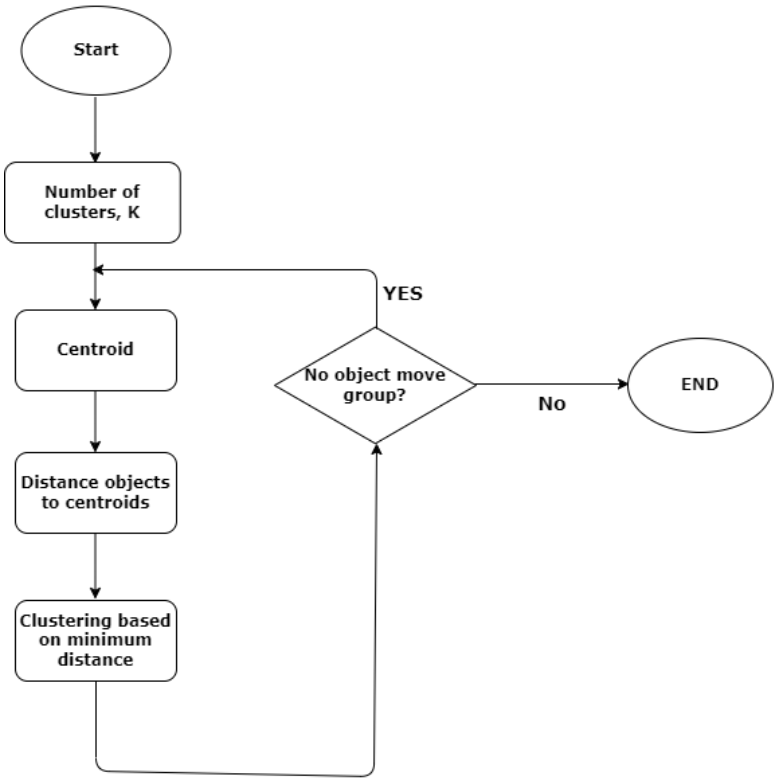


Figure 2.5 Flowchart of a K-means clustering algorithm

2.9.2 DBSCAN clustering

DBSCAN is a clustering algorithm used in unsupervised ML. It is designed to identify clusters of data points that are densely packed together in a high-dimensional space. The algorithm works by defining a neighborhood around each data point based on a radius, called epsilon(ε), and a minimum number of points, called min_samples, that must be present in that neighborhood to form a dense region. Points that are not part of any dense region are labeled as noise. An advantage of this algorithm is that can be applied to datasets that contain noise points because this algorithm is insensitive to noise points. This algorithm identifies these noise points and excludes them from clustering results [33]. However, it can be sensitive to the choice of hyperparameters mentioned previously, which can affect the clustering results. The DBSCAN method should be used to detect correlations and structures in data that are challenging to find manually but may be pertinent and valuable to identify patterns and forecast trends [34].

The algorithm starts by selecting an arbitrary data point and finding all other points that are within its epsilon radius. If the number of points in this neighborhood is greater than or equal to the minimum sample threshold, then a new cluster is formed. The algorithm then continues to explore the points in this cluster, expanding it until all points within the cluster have been identified. Once a cluster has been identified, the algorithm moves on to the next unexplored data point and repeats the process. Figure 2.6 how this algorithm clusters data points.

In comparison to K-means, DBSCAN is better suited for datasets with complex shapes and varying densities, where the number of clusters is not known prior to clustering. It can also handle outliers well since it labels them as noise. K-means, on the other hand, is more appropriate for datasets with a fixed number of clusters and where the clusters are roughly spherical and equally sized as depicted in Figure 2.7.

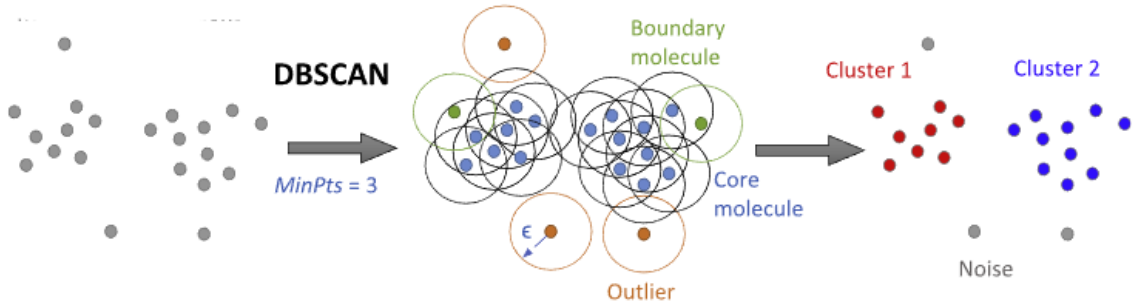


Figure 2.6 DBSCAN algorithm [35]

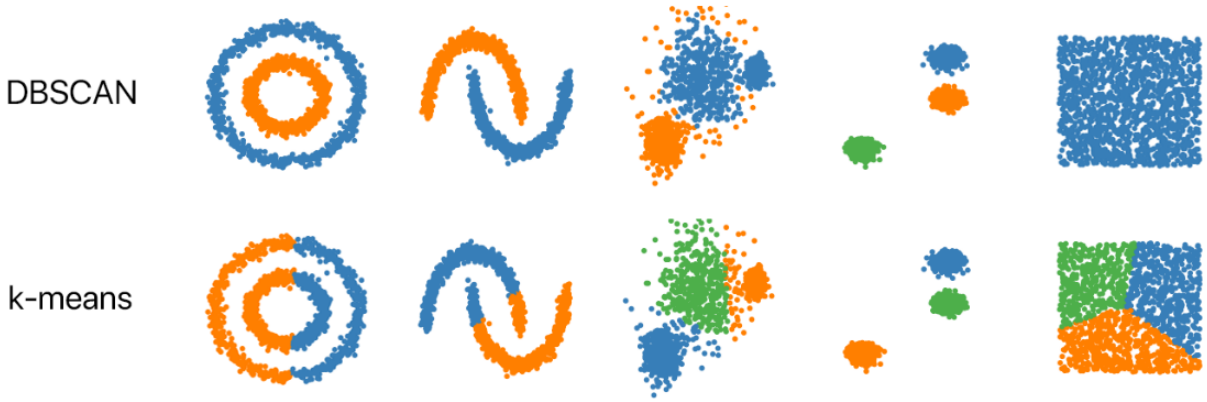


Figure 2.7 DBSCAN algorithm compared to K-means algorithm [36]

3. DATA ACQUISITION
3.1 Introduction

Data acquisition refers to the process of acquiring data from the environment. The process of data acquisition facilitates obtaining specific knowledge about the environment being studied [37]. Data collection, cleaning, processing, and transformation into a usable version are all steps in the crucial data acquisition process of data analysis and Machine learning model creation. Making informed choices requires accurate and reliable data, which can only be obtained through high-quality data acquisition. In this part, we describe how the temperature, CO_2 levels and ventilation consumption dataset from a school in Tartu was collected and pre-processed for model creation. We outline the procedures used to prepare, process, and convert the data into an analytically usable format.

3.2 Data source

The data used in this study was obtained from a recently renovated educational building in Tartu, Estonia. Ventilation units with balanced air flow and heat recovery systems were installed in the building during the renovation. The dataset consisted of:

- CO_2 sensor data for 57 classes
- Temperature sensor data for the 57 classes
- Cumulative ventilation data for 11 ventilation units in the building.

The period of data collection spanned from 01.09.2020, to 31.01.2022 and the frequency of sampling was at 5-minute intervals for 24-hrs per day including weekends and holidays. The dataset per class contained 154.885 records. The CO_2 levels was measured using commercial indoor air quality NDIR sensors SMT-IAQ3 visible in Appendix 1 which has an accuracy of +/-30ppm at 25°C and an operating range of 0...2000ppm. According to the manual [38], the sensor employs an advanced learning self-calibration feature that takes place over the course of eight days. The data used for this work included 3 classes with volumes 153 m^3, 168 m^3 and 200 m^3 using ventilations units labelled SV06, SV05 and SV07 respectively.

3.3 Data cleaning

The first step in data processing was cleaning the data. Data cleaning is one of the most important processes in data analysis and the first step in any machine learning endeavor. It is a crucial stage in making sure the dataset is free of erroneous data. For the creation of high-performing, accurate Machine Learning (ML) applications, the availability of high-quality data is a requirement. But in practice, data is rarely clean because of erroneous inputs from manual data curation or inevitabilities in automated data collection or generation processes [39]. Erroneous data can prevent algorithms from finding patterns, while cleaned data guarantees consistency in model training for the best outcomes. According to recent research by Forbes, data scientists typically spend 80% of project effort cleaning data [40]. This demonstrates the value of data cleaning and how time-consuming it is. Prior to conducting further research, it was necessary to resolve the missing values, outliers, and discrepancies in the study's raw data. There was also the problem of no noticeable changes in the readings from the CO_2 and temperature sensors when the building was completely unoccupied, for example, on weekends and vacations, as CO_2 and temperature are proportionate to human activity in a space. The steps taken to clean the data are discussed below.

3.3.1 Removing missing (null) values

Missing data happens in practically all studies, including those that are carefully planned and managed. The statistical power of a study can be decreased by missing data, which can also lead to erroneous estimates and inaccurate conclusions [41]. This was the case with the data for this study at the initial stage. Missing data can be handled by one of the many approaches which include simply omitting those cases which have missing values and analyzing the remaining data. This is called listwise or case deletion, it is by far the most common approach and is what is used in this study as can be seen from Figure 3.1. Another approach to handling missing values is pairwise deletion which involves excluding observations that have missing data on any variable involved in an analysis, only for the specific pairwise comparisons that are being performed, it is also known as case deletion and results in preservation of more information than the former.

Mean substitution is a method of handling missing values where the missing data value for a variable is replaced with the variable's mean value. The “Regression imputation” method handles missing data by estimation. The missing values are estimated by regressing that variable on other variables in the dataset that are related to it. The regression model is used to predict the missing values based on the values of the other variables. The “last observation carried forward” method is another common approach of handling missing data in time series data. It involves replacing a missing value with the last observed value. It can be done forward or backward. It was also used in this study to fill missing values in the ventilation dataset. Mean substitution involves replacing missing values with the mean values of a variable in the dataset.

	timestamp	köök (kWh)	SV01 (kWh)	SV02 (kWh)	SV03 (kWh)	SV04 (kWh)	SV05 (kWh)
0	2020-07-06 17:00:00	NaN	NaN	NaN	NaN	NaN	NaN
1	2020-07-06 18:00:00	NaN	NaN	NaN	NaN	NaN	NaN
2	2020-07-06 19:00:00	NaN	NaN	NaN	NaN	NaN	NaN
3	2020-07-06 20:00:00	NaN	NaN	NaN	NaN	NaN	NaN
4	2020-07-06 21:00:00	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...
13705	2022-01-31 19:00:00	79863.31	2399.39	3649.83	13044.88	8286.78	21454.66
13706	2022-01-31 20:00:00	79864.45	2399.44	3650.22	13044.97	8286.84	21456.79
13707	2022-01-31 21:00:00	79865.47	2399.48	3650.57	13045.04	8286.91	21456.86
13708	2022-01-31 22:00:00	79866.66	2399.52	3650.97	13045.12	8286.97	21456.93
13709	2022-01-31 23:00:00	79867.78	2399.56	3651.28	13045.21	8287.04	21456.99

Figure 3.1 Raw data with missing values (NaN) which have to be eliminated

timestamp	köök (kWh)	SV01 (kWh)	SV02 (kWh)	SV03 (kWh)	SV04 (kWh)	SV05 (kWh)
2020-09-01 00:00:00	2501.65	838.01	901.85	1504.89	2466.85	5587.81
2020-09-01 01:00:00	2502.46	838.55	902.49	1504.94	2466.91	5591.18
2020-09-01 02:00:00	2503.17	839.07	903.07	1504.98	2466.97	5594.19
2020-09-01 03:00:00	2503.90	839.58	903.71	1505.03	2467.04	5597.52
2020-09-01 04:00:00	2504.66	840.11	904.35	1505.07	2467.11	5600.64
...	...	...	...	...	...	...
2022-01-31 19:00:00	79863.31	2399.39	3649.83	13044.88	8286.78	21454.66
2022-01-31 20:00:00	79864.45	2399.44	3650.22	13044.97	8286.84	21456.79
2022-01-31 21:00:00	79865.47	2399.48	3650.57	13045.04	8286.91	21456.86
2022-01-31 22:00:00	79866.66	2399.52	3650.97	13045.12	8286.97	21456.93
2022-01-31 23:00:00	79867.78	2399.56	3651.28	13045.21	8287.04	21456.99

Figure 3.2 Resulting data after eliminating missing values using case deletion method

3.3.2 Removing weekends and holidays

Since the data was collected over a period of two years, it included weekends and holidays. However, due to decreased occupancy levels on weekends and holidays, the CO_2 and temperature levels were noticeably lower. To remove noise and straight lines from the dataset, weekends and holidays were taken out. It is important to remove weekends and holidays from the CO_2 data when building an ML model to predict occupancy because the occupancy patterns on weekends and holidays can be significantly different from those on weekdays. We can make sure that the model is trained on data that is typical of weekday occupancy patterns, which are probably more regular and predictable, by excluding weekends and holidays. This may help increase the model's predicted occupancy's accuracy. The visualization shown in Figure 3.3 below supports the claim that CO_2 level is proportional to occupancy and can be used to predict building occupancy as there is no increase in the observed parameters' levels during the weekend where CO_2 levels on the y-axis remain constant all through the time of the day on the x-axis.

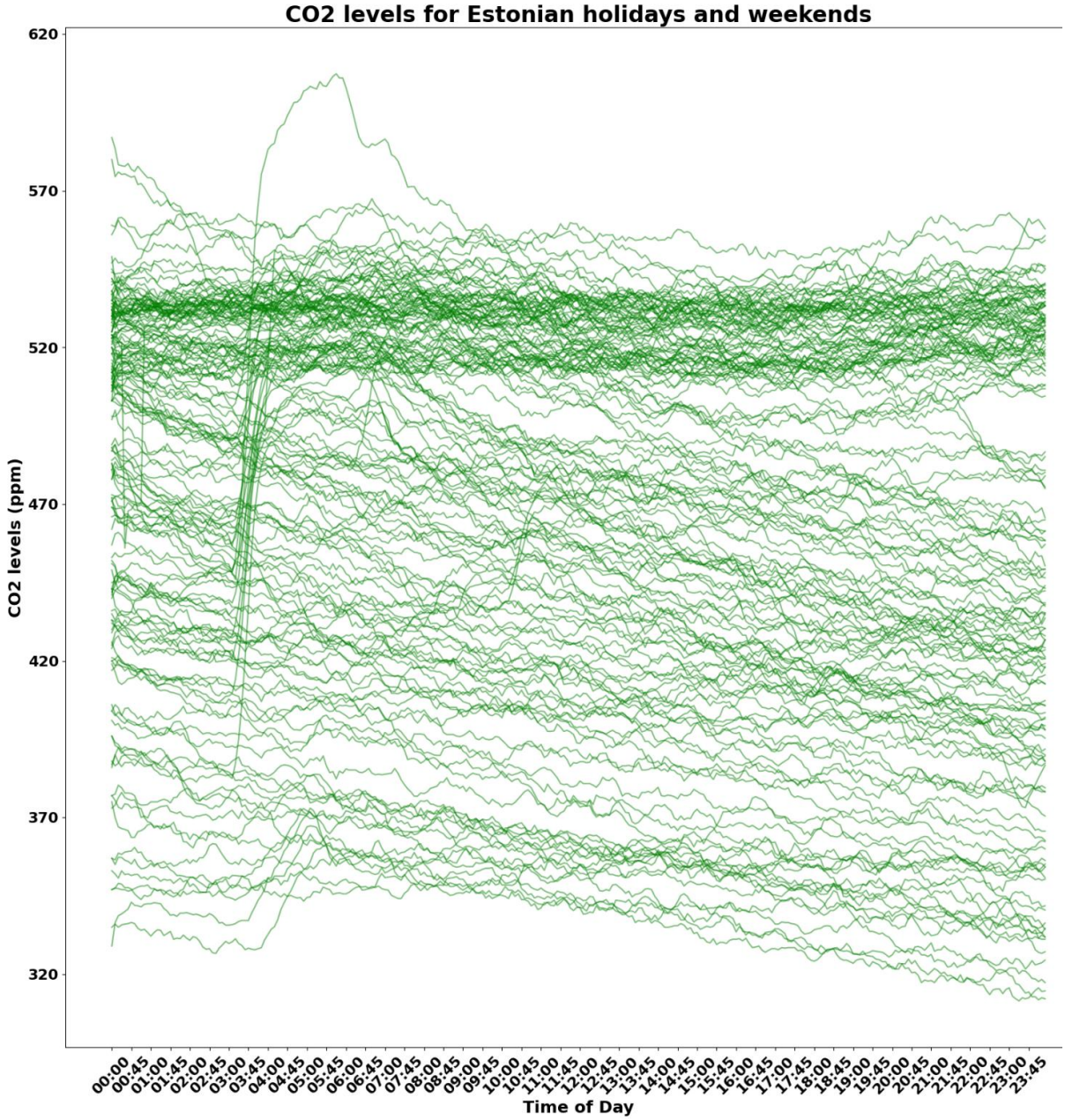


Figure 3.3 Noisy and straight plots caused by weekend and holiday data validating the claim of CO_2 level being proportional to occupancy

3.3.3 Fixing baselines to accommodate sensor calibration

Fixing the baselines of CO_2 data refers to a process involving adjusting the CO_2 concentration measurements to a reference point, often to a value that corresponds to the normal or anticipated range of CO_2 in the environment. This is done to focus on the more significant changes in CO_2 levels that signal the occupancy state of the space and to reduce the impact of the natural variations in CO_2 concentration caused by things like human breathing and ventilation. Since occupancy patterns and CO_2 levels have a high correlation, fixing the baselines of the CO_2 data is crucial when utilizing it to estimate occupancy. By fixing these baselines, the model will be trained to distinguish between CO_2 levels that are attributable to occupancy and those that are not. Baselines relate to the CO_2 levels when the space is empty. Instead of discriminating between increases in CO_2 due to occupancy and increases due to other factors, the model may learn to forecast occupancy based only on CO_2 levels if the baselines are not constant. This may result in incorrect predictions and a model that is not applicable in real-world situations. By adjusting the baselines, the model can be trained to recognize occupancy patterns more accurately, improving estimation accuracy. Fixing the baselines, in other words, means reducing the noise in the CO_2 data brought on by variables unrelated to occupancy, allowing the ML model to concentrate on the patterns and trends more pertinent to predicting occupancy. This is frequently accomplished by using a statistical filter to eliminate noise or by deducting a baseline value from the raw CO_2 data. Figures 3.4 and 3.5 show a comparison of the same data before and after fixing the baselines accurately. Since the accepted CO_2 level for unoccupied spaces in 400ppm, this value was used when fixing baselines using a 288-window period which corresponds to (24hours * 60min) / 5-minute intervals.

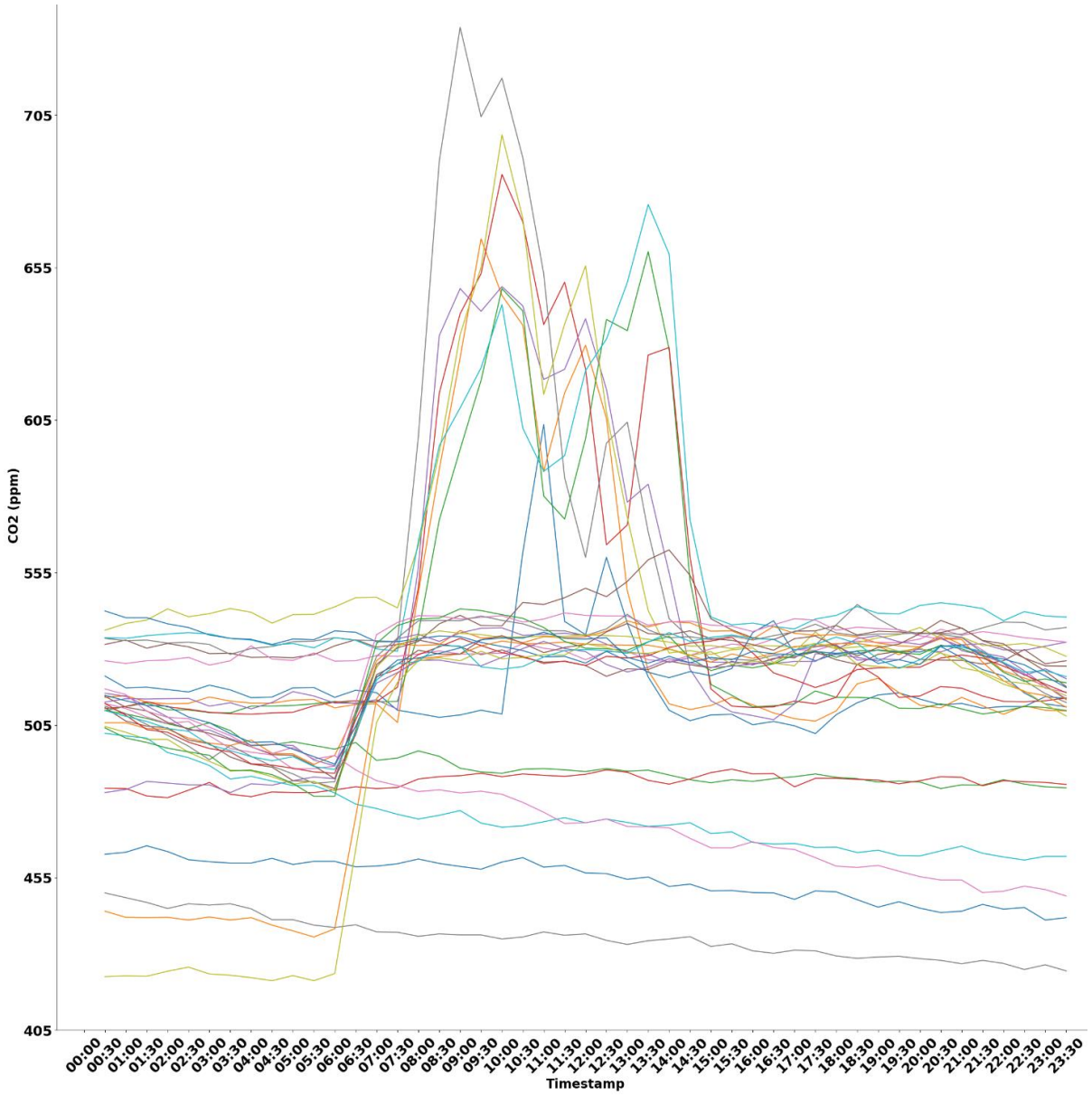


Figure 3.4 CO_2 level against the time of day plot showing data before fixing CO_2 baseline for calibration

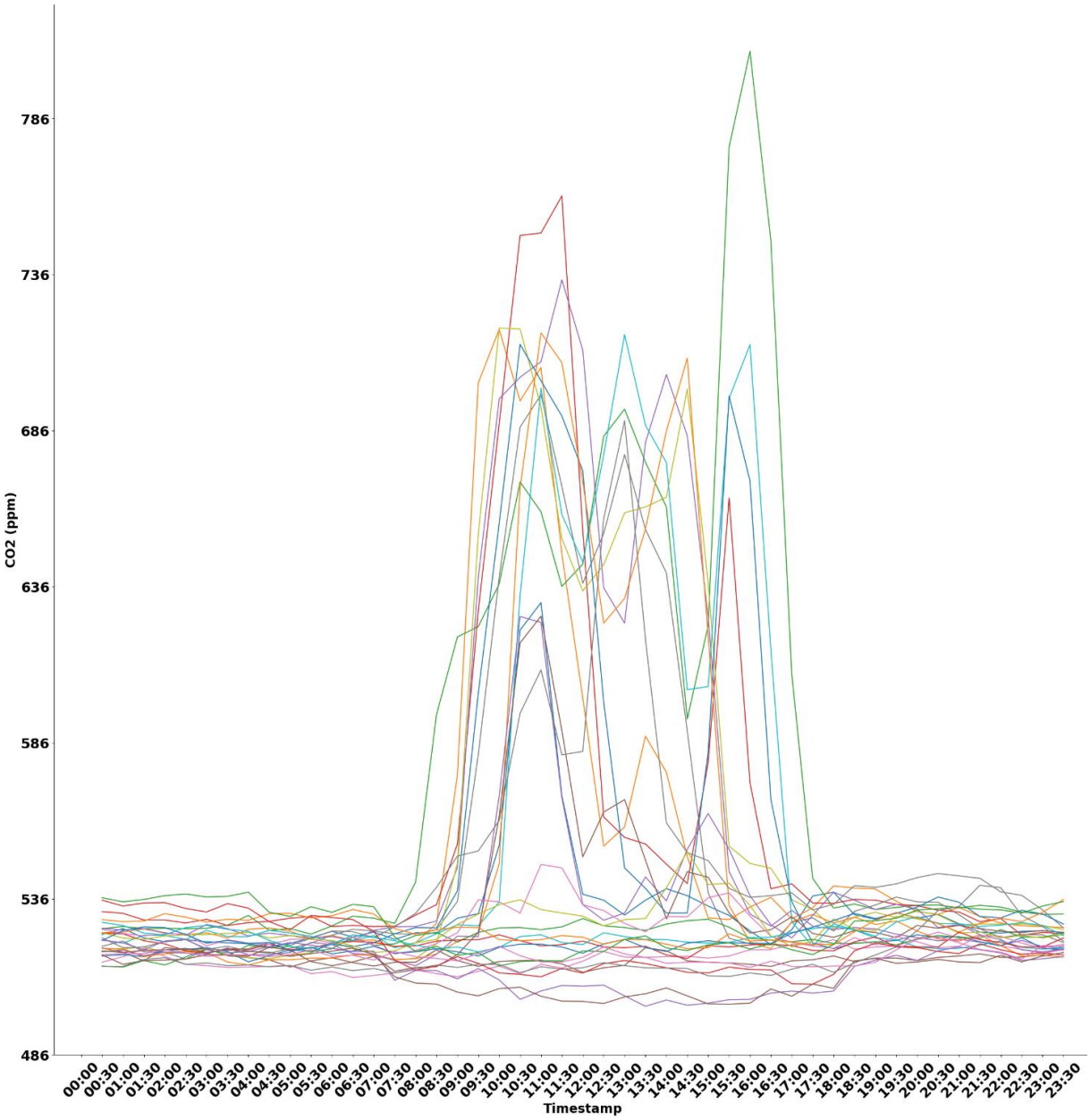


Figure 3.5 CO_2 level against the time of day plot showing cleaner and useable data after fixing CO_2 baseline

3.3.4 Calculation of Distinct Consumption Data from Cumulative Consumption

To accurately match the CO_2 data interval, the cumulative ventilation consumption statistics were approximated into distinct consumption values per 5-minute interval by resampling the 1-hour interval using the mean approach, as shown in Figure 3.6. To achieve this, an empty pandas series with a specified data type was created using a function that calculated the difference between succeeding values repeatedly in a loop. The remaining values were then filled using a backfill technique. The resulting series provided a more granular view of the ventilation consumption, allowing for a more precise analysis of the relationship between ventilation and indoor air quality. Figure 3.7 displays the final outcome of this process. However, future works can explore more sophisticated methods of approximating cumulative values and resampling, which could potentially improve the accuracy of the analysis.

timestamp	SV05 (kWh)
2020-09-01 00:00:00	5587.81
2020-09-01 01:00:00	5591.18
2020-09-01 02:00:00	5594.19
2020-09-01 03:00:00	5597.52
2020-09-01 04:00:00	5600.64
...	...
2022-01-31 19:00:00	21454.66
2022-01-31 20:00:00	21456.79
2022-01-31 21:00:00	21456.86
2022-01-31 22:00:00	21456.93
2022-01-31 23:00:00	21456.99

(a)

timestamp	SV05 (kWh)
2020-09-01 00:00:00	5587.810
2020-09-01 00:05:00	5588.091
2020-09-01 00:10:00	5588.372
2020-09-01 00:15:00	5588.652
2020-09-01 00:20:00	5588.933
...	...
2022-01-31 23:35:00	21456.990
2022-01-31 23:40:00	21456.990
2022-01-31 23:45:00	21456.990
2022-01-31 23:50:00	21456.990
2022-01-31 23:55:00	21456.990

(b)

Figure 3.6 Cumulative ventilation consumption data before and after re-sampling. (a) shows consumption with 1-hr timestamps and (b) shows the resulting consumption levels with 5-minutes timestamps

timestamp	SV05 (kWh)
2020-09-01 00:05:00	0.281
2020-09-01 00:10:00	0.281
2020-09-01 00:15:00	0.280
2020-09-01 00:20:00	0.281
2020-09-01 00:25:00	0.281
2020-09-01 00:30:00	0.281
2020-09-01 00:35:00	0.281
2020-09-01 00:40:00	0.281
2020-09-01 00:45:00	0.281
2020-09-01 00:50:00	0.280
2020-09-01 00:55:00	0.281
2020-09-01 01:00:00	0.281

Figure 3.7 Ventilation data after calculating distinct values

3.3.5 Grouping the CO_2 data based on ventilation levels

To gain a better understanding of the ventilation data and extract significant insights, a visualization technique was applied. Specifically, the CO_2 and temperature levels were grouped by week to provide a comprehensive view of the data and allow for the identification of weekly patterns in CO_2 levels. To further investigate the impact of increasing population on CO_2 levels, the CO_2 data was grouped according to ventilation levels, and these grouped weeks were subsequently clustered based on weeks where the ventilation was maintained at a specific level. The purpose of this grouping was to generate a dataset with sufficient amount of comparable data points that could be utilized for model creation, as the quantity of related data points also influences the model's accuracy. Through this approach, it was possible to identify CO_2 levels at peak occupancy and low occupancy using percentiles, providing valuable insights into the relationship between ventilation and CO_2 levels.

3.3.6 Eliminating outliers

Outliers are observations or measurements that are unusually little or large compared to the bulk of the observations, and hence are suspicious [42]. They are data points that do not fit the series' historical trend or regular pattern of change. These observations are troubling because they might not be the result of the process being examined or might not accurately represent the trend under observation and as a result, a prediction model may learn erroneous information when the data it is fed is contaminated by these aberrant values [43]. Even a few outliers can sometimes be enough to skew the results of the group by altering the mean performance or by increasing variability unnecessarily.

In the dataset under consideration, there existed some outliers that needed to be eliminated after grouping the data by ventilation levels. A representation of one of the outliers available in the dataset is shown in Figures 3.8 and 3.9 before and after outliers were eliminated, respectively.

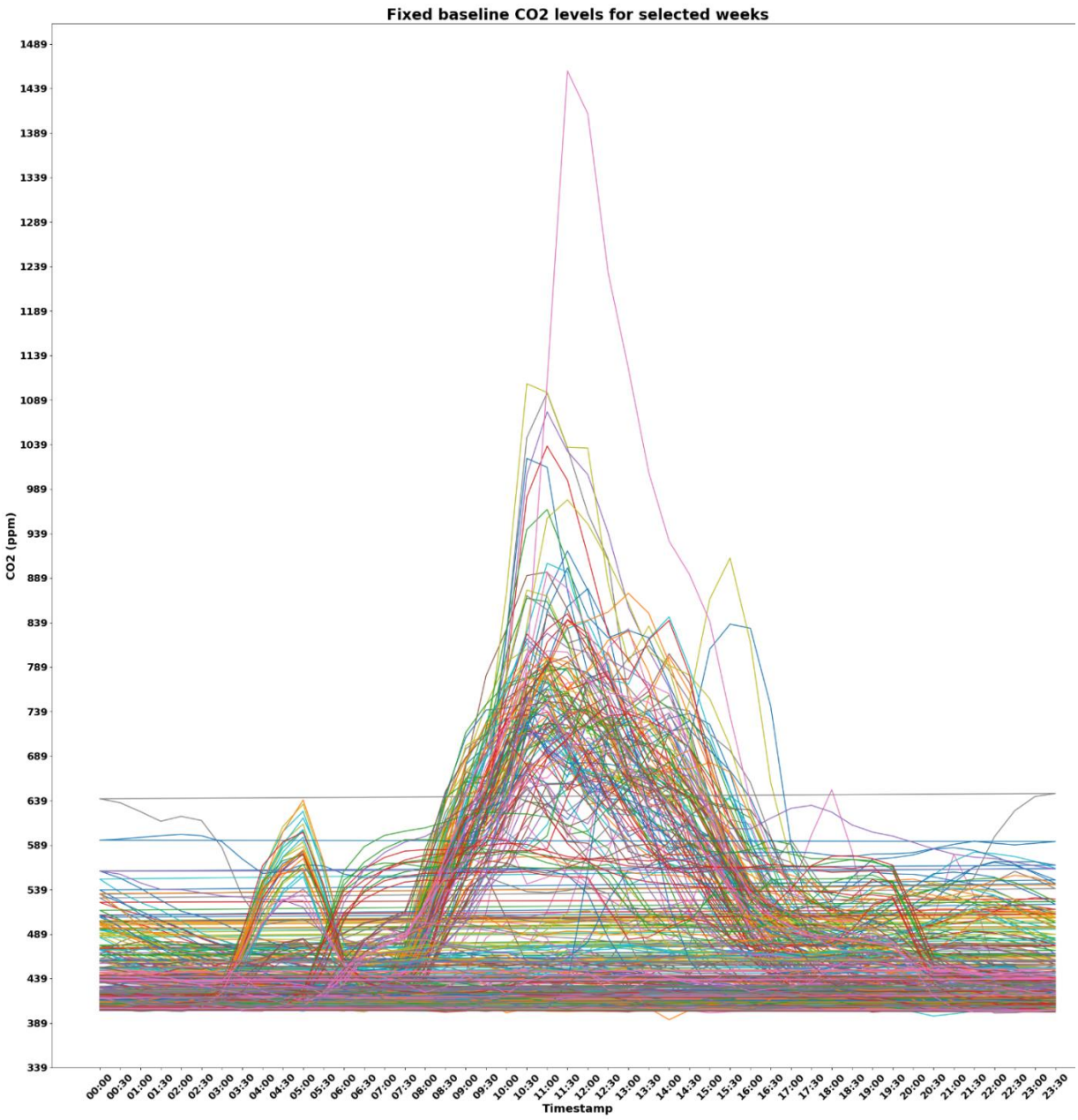


Figure 3.8 CO_2 data plot including outliers

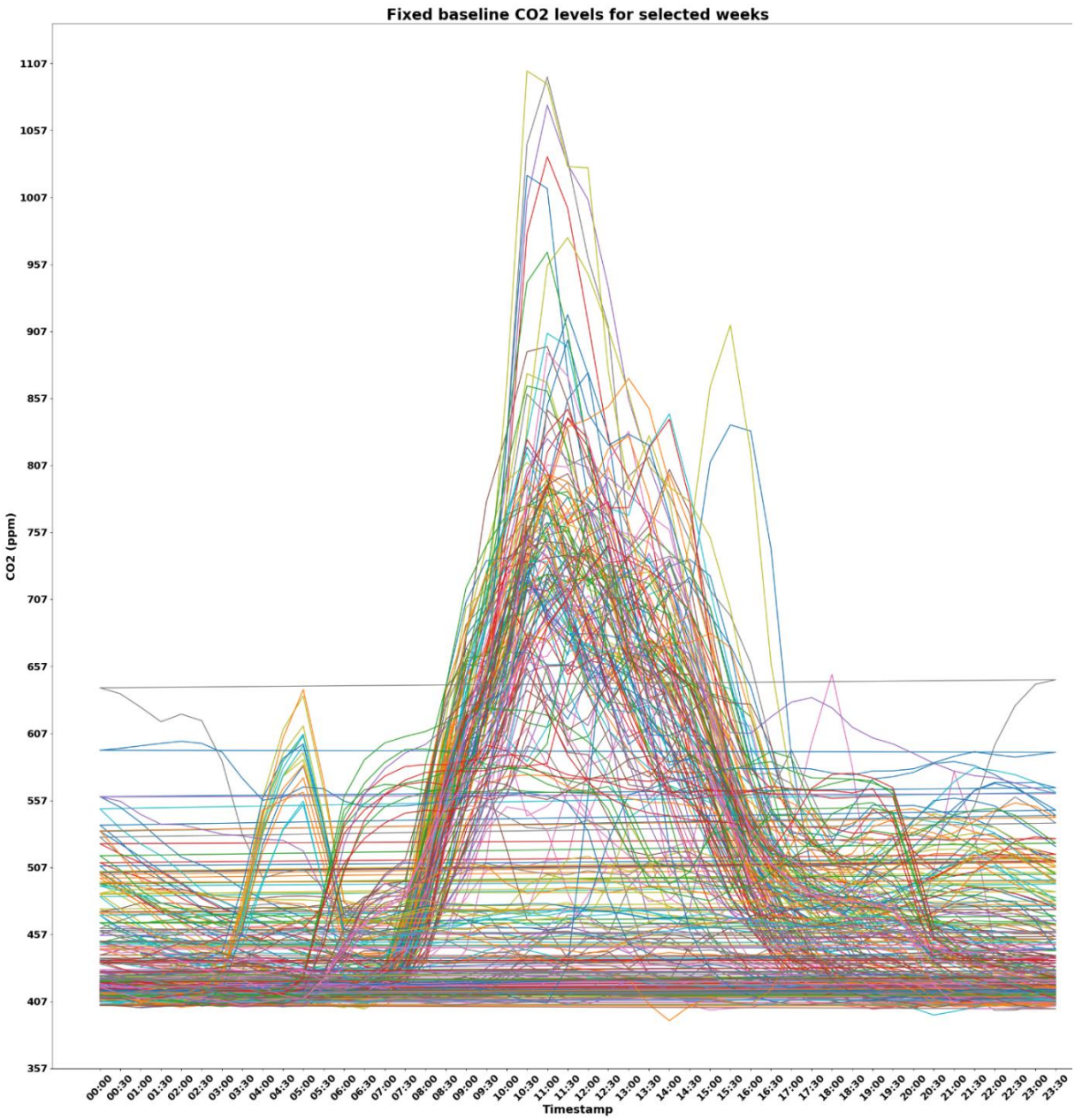


Figure 3.9 CO_2 data plot after eliminating outliers

4. ANALYSIS AND TEST RESULTS
4.1 Overview of the analysis

This chapter presents a detailed analysis of the occupancy estimation methods used in this study, utilizing the preprocessed data obtained in Chapter 3, and it also describes the validation of the applied method. The chapter is structured into three main sections, in accordance with the framework established in Chapter 2. The first section provides a comprehensive explanation of the method adopted in this study. The second section outlines the process of selecting the appropriate machine learning algorithm to be used for occupancy estimation based on the selected method. The final section focuses on the validation of the methodology using the CO_2 mass balance equation for gaseous spaces and mixing. Through these sections, this chapter aims to provide a clear and detailed understanding of the occupancy estimation methods utilized in this study, and their validity in estimating the number of occupants in the selected space.

4.2 Methodology

The method used in this study is the estimation of occupancy using CO_2 and temperature data based on the historical energy consumption data for the ventilation unit installed in the building. Since the ventilation units only run during a specific time when the building is occupied, the main meter installed in the building under consideration can be used to confirm that there are occupants there, but this information is insufficient to accurately estimate the occupancy level. The historical consumption data can be utilized in such a way that we take into account days when the units were working at the same power level, keeping in mind that if the unit works at the same level, we can then analyze the build-up of CO_2, keeping in mind that the more people present, the higher the CO_2, and these can be clustered together as opposed to using CO_2 information for various times when the ventilation units work at different power levels. When air is pumped into the space and removed at different rates, we cannot precisely say that it is fully occupied at a certain CO_2 level and group them together. This presumption is only valid if the air is entering and leaving the space at the same rate for all levels of CO_2.

This method can then be verified by utilizing the CO_2 mass balance equation to determine the population at a specific moment and demonstrating how full occupancy for one flow rate differs from that at another. Three factors led to the choice of this approach. These are:

- CO_2 and temperature sensors are readily available in buildings and hence no extra amenities will need to be installed which will increase the cost for all schools in the building under consideration. Every classroom for students and teachers has been fitted with CO_2 and Temperature sensors.
- The privacy concerns attached to obtaining ground truth with other known methods. As the occupants of the building in question are underaged children, care must be taken when trying to obtain ground truth for validation of any methodology adopted for occupancy estimation.
- Clusterization Unsupervised ML algorithms due to the type and quantity of data. A cluster is a set of core samples that are close to each other (measured by some distance metrics) and a set of non-core samples close by that are not considered core samples themselves. The goal of clusterization is to identify patterns or similarities in the data without any prior knowledge or labelling of the data. Clustering algorithms usually use distance or similarity measures to group data points together based on their features or attributes. The goal of classification is to learn a model that can accurately predict the label or class of new data points based on the training data. The clusterization algorithms sampled in this study include K-means, GMM and DBSCAN.

4.2.1 Grouping based on ventilation consumption level

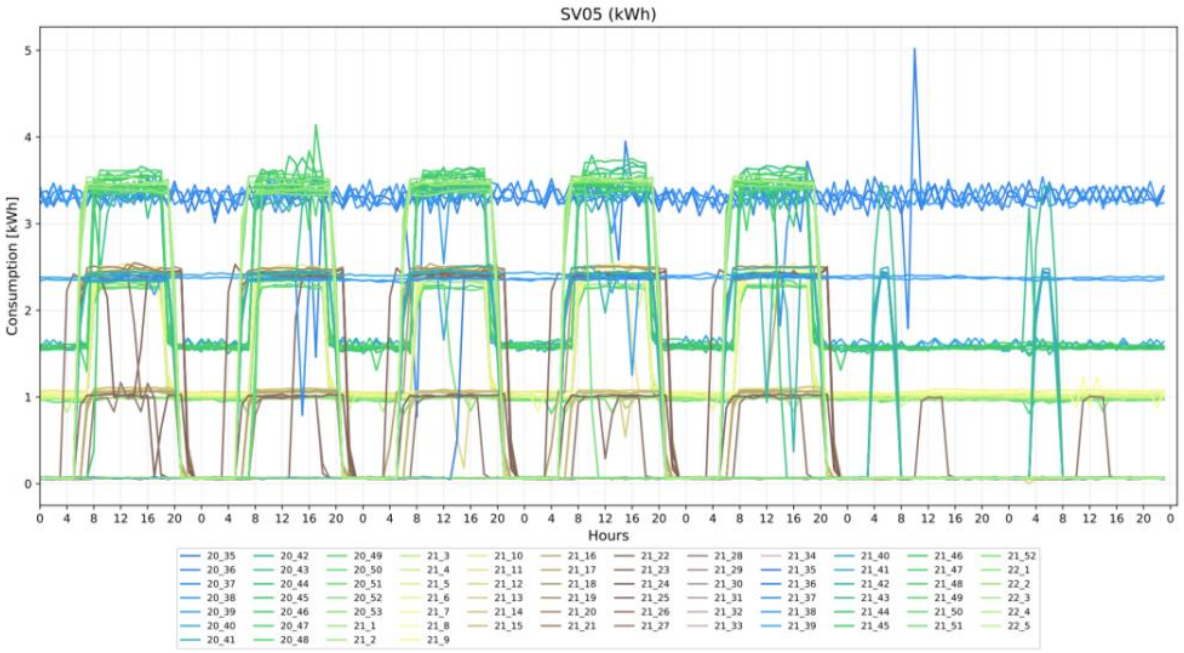
As explained in Section 3.3.5, the preprocessed CO_2 and Temperature data was grouped based on the ventilation unit consumption level. At a certain energy consumption level, the power and flowrate are the same and the clustering can then be carried out. The Figure 4.1 shows the grouping of the three ventilation units under consideration using their energy consumption data to show the levels at which the unit was working during the data collection period.

This grouping what then used to group the CO_2 and Temperature data of these days together. The x-axis shows the hours for one week. 0 to 24 for a 24-hr period for 5 days (Monday, Tuesday, Wednesday, Thursday, and Friday) as seen below.

This represents one week data on the x-axis and consumption data on the y-axis. On weeks where the units work at the same level, the CO_2 and Temperature data was grouped together to make a dataset. For example weeks 20_42, 20_43, 20_44, etc, were grouped together for ventilation unit SV05. The weeks are labelled year_week number since there are 52 weeks in the year.

Upon grouping, the dataset with enough datapoint to be fed into the into a ML algorithm was selected. This was only 1 dataset per class as shown from Figure 4.2 which shows a dataset with enough datapoint suitable for ML as seen in and one dataset with very few datapoint as seen in Figure 4.3. The units have a layout in Appendix 2 and are labelled as follows:

- Class 1 uses SV06
- Class 2 uses SV05
- Class 3 uses SV07



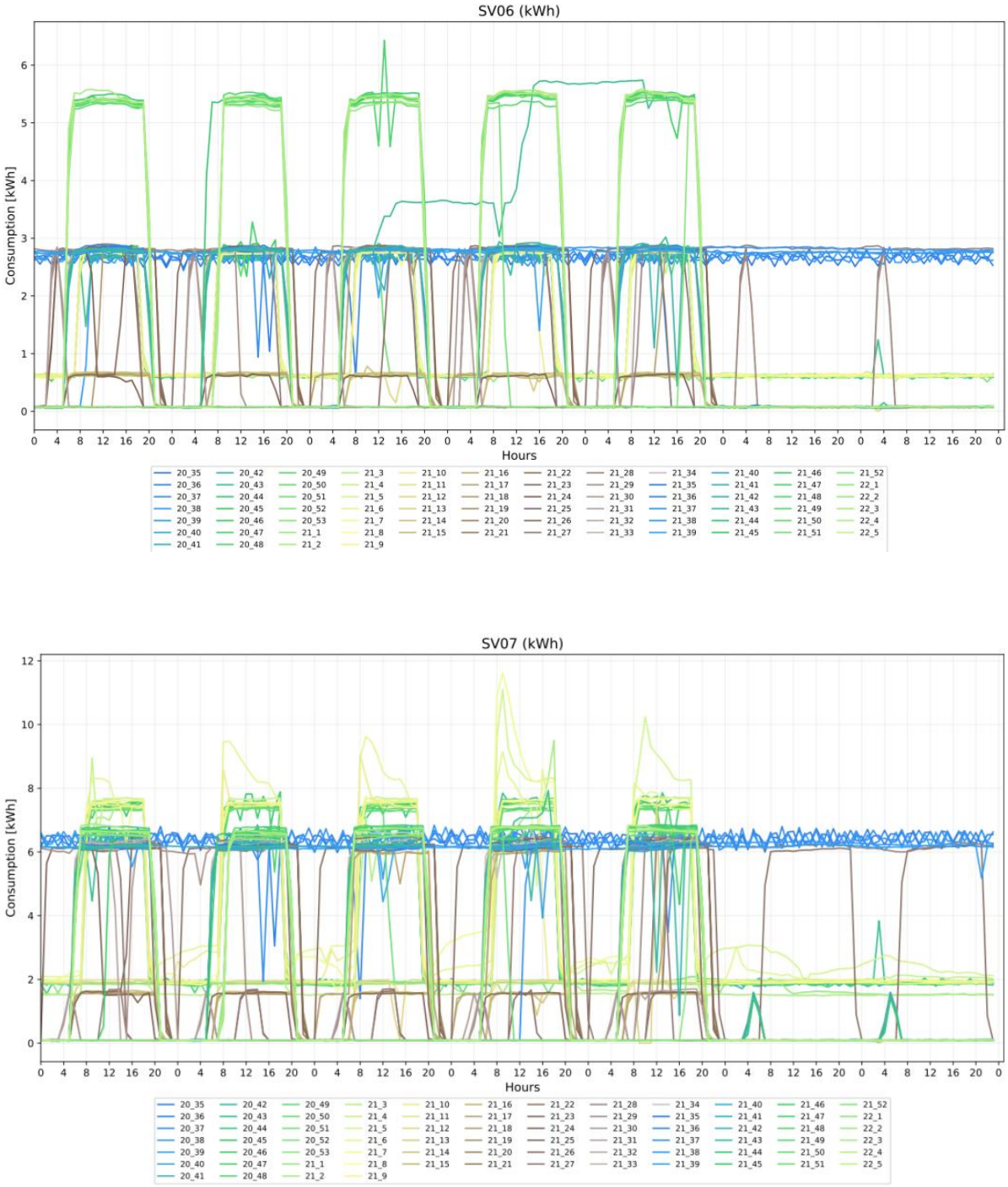


Figure 4.1 Ventilation consumption levels per week for the three Ventilation units under consideration

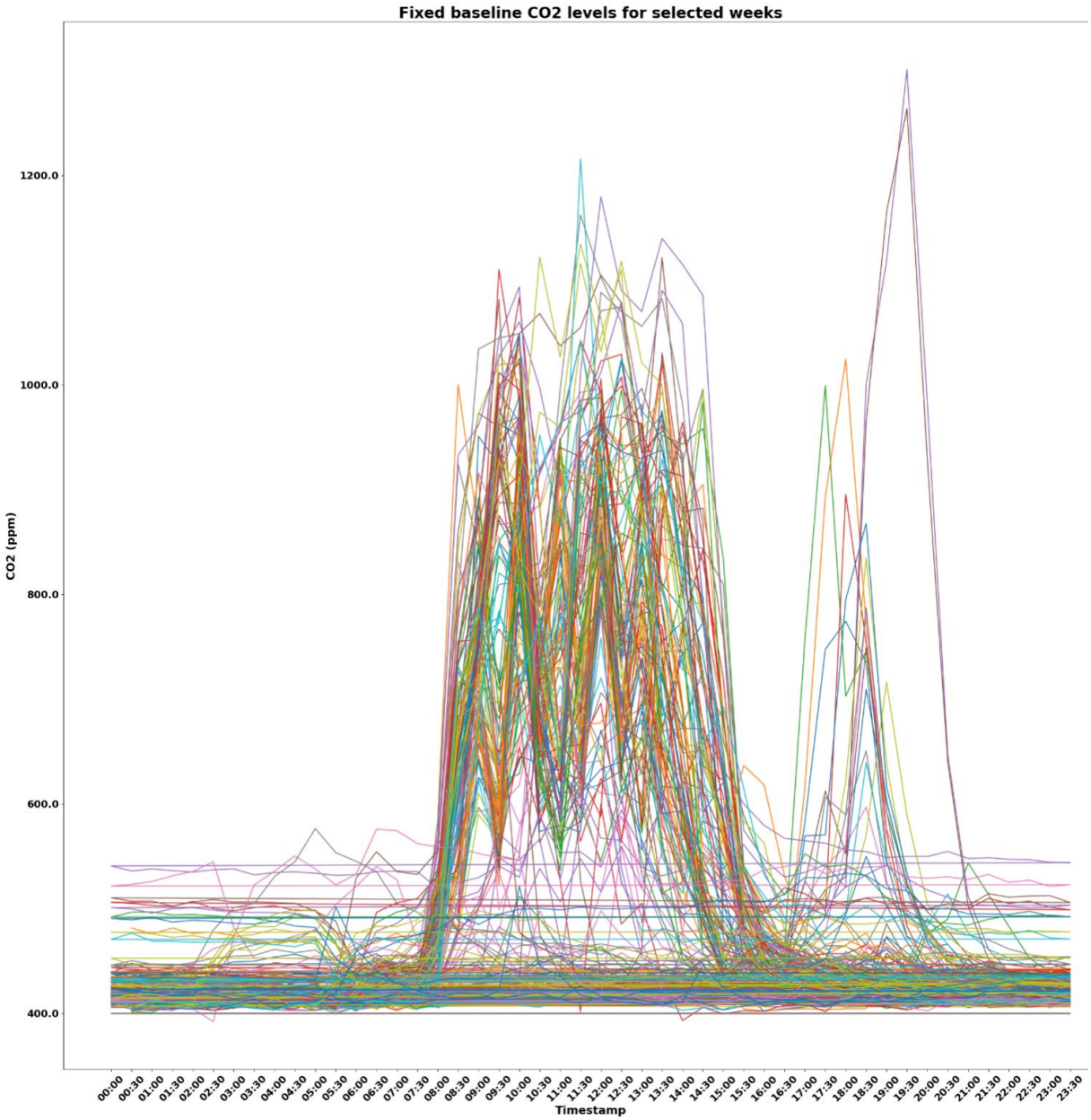


Figure 4.2 Grouped data with required datapoints for ML

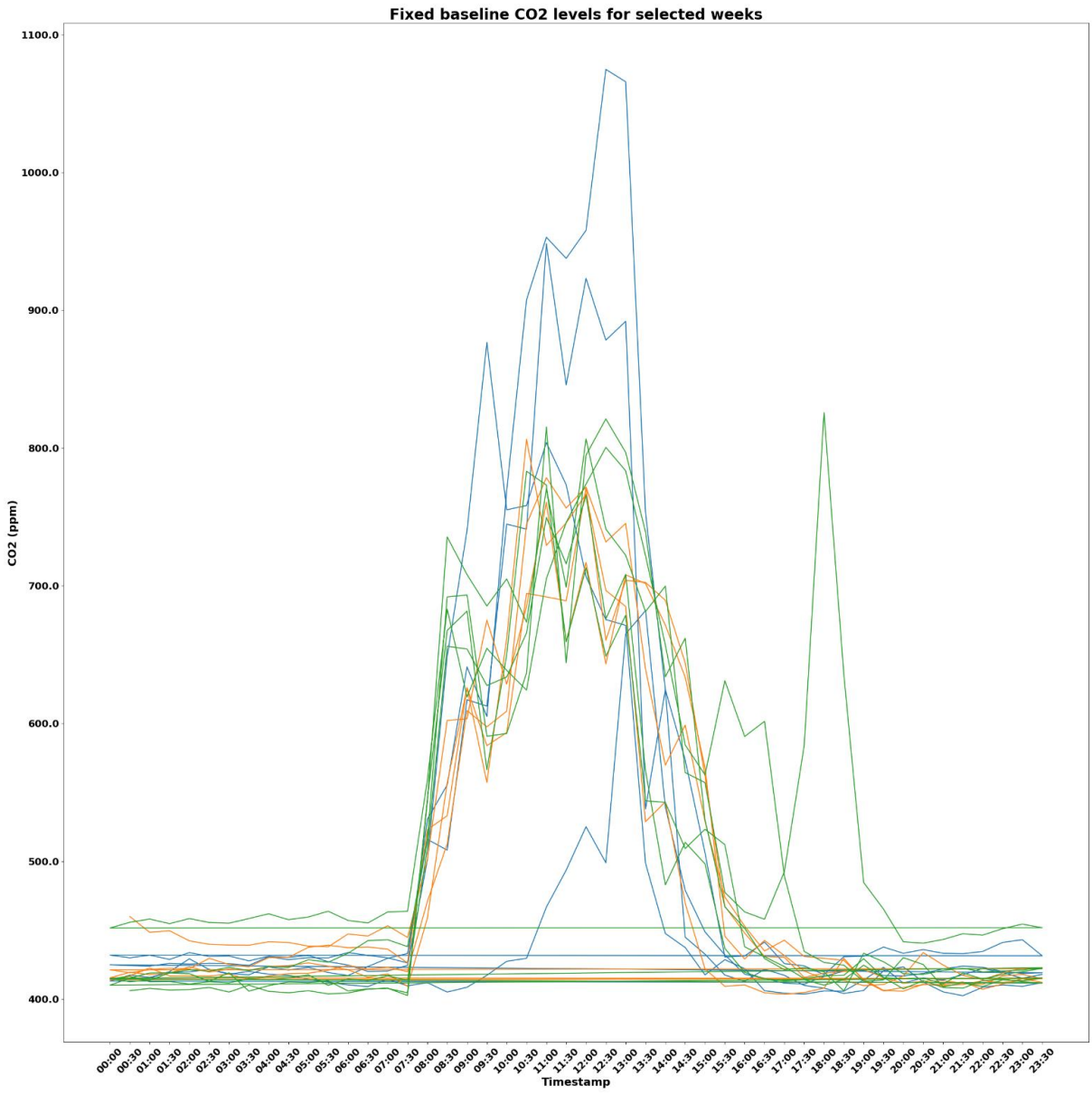


Figure 4.3 Grouped data with insufficient datapoint

4.3 Selected ML algorithm

In the absence of ground truth, the ML algorithm to be used will be an unsupervised learning method for reasons explained in Section 2.7.2. Three algorithms were implemented for all 3 classes under consideration and the best algorithm that properly clusters the data was selected. The correct number of clusters for the data was arrived at using the elbow method. Figure 4.4 shows the elbow method applied to each useable dataset for all three classes under consideration. In the Elbow method, the number of clusters (K) is varied, then the WCSS is calculated for each value of K. The plot of the WCSS with the K value resembles an elbow. The idea is that, as the WCSS value drops, the number of clusters rises. When the graph is examined, an elbow is observed due to the rapid change between WCSS and K. At a point, the graph moves parallel to the x-axis. The ideal K value is the one that corresponds to this point [44]. Figure 4.4 shows this representation. From Figure 4.5, the ideal number of clusters for all dataset under consideration is three (3) as indicated by the arrow. The clusters will be labelled as "Low occupancy", "Medium occupancy", and "Full occupancy".

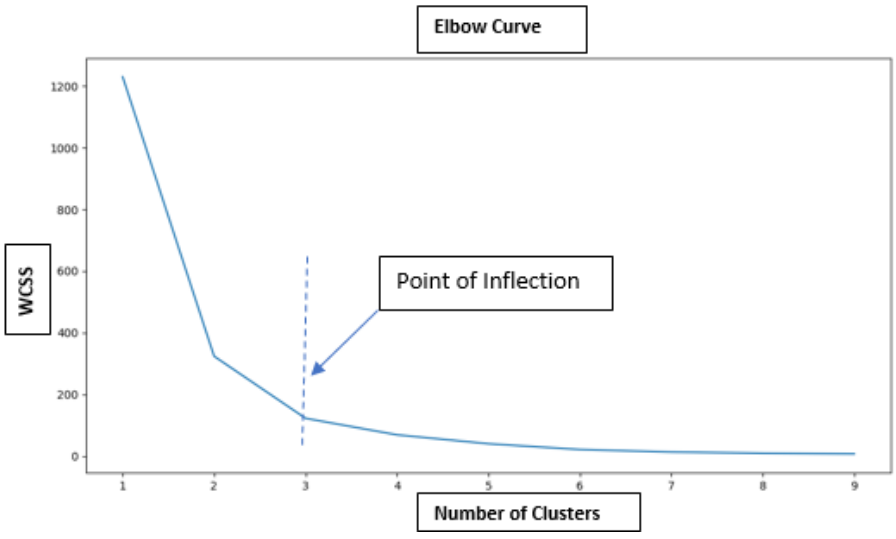


Figure 4.4 Sample elbow plot [45]

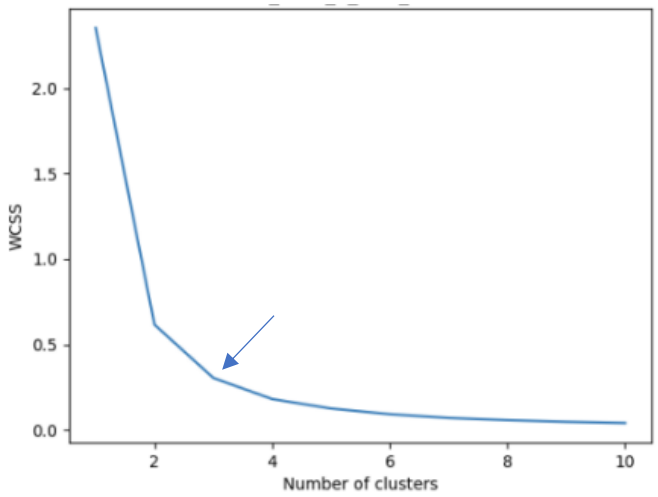


Figure 4.5 Elbow plots for this work's dataset with point of inflection of "3"

Figures 4.6 and 4.7 depict the daily and weekly fluctuations in CO_2 level, with each class session lasting 45 minutes and starting at 8:00 am. These fluctuations are reflected in the rise and fall of the plot for the first 45 minutes of each day, providing a visual representation of occupancy level as students enter and leave the classroom.

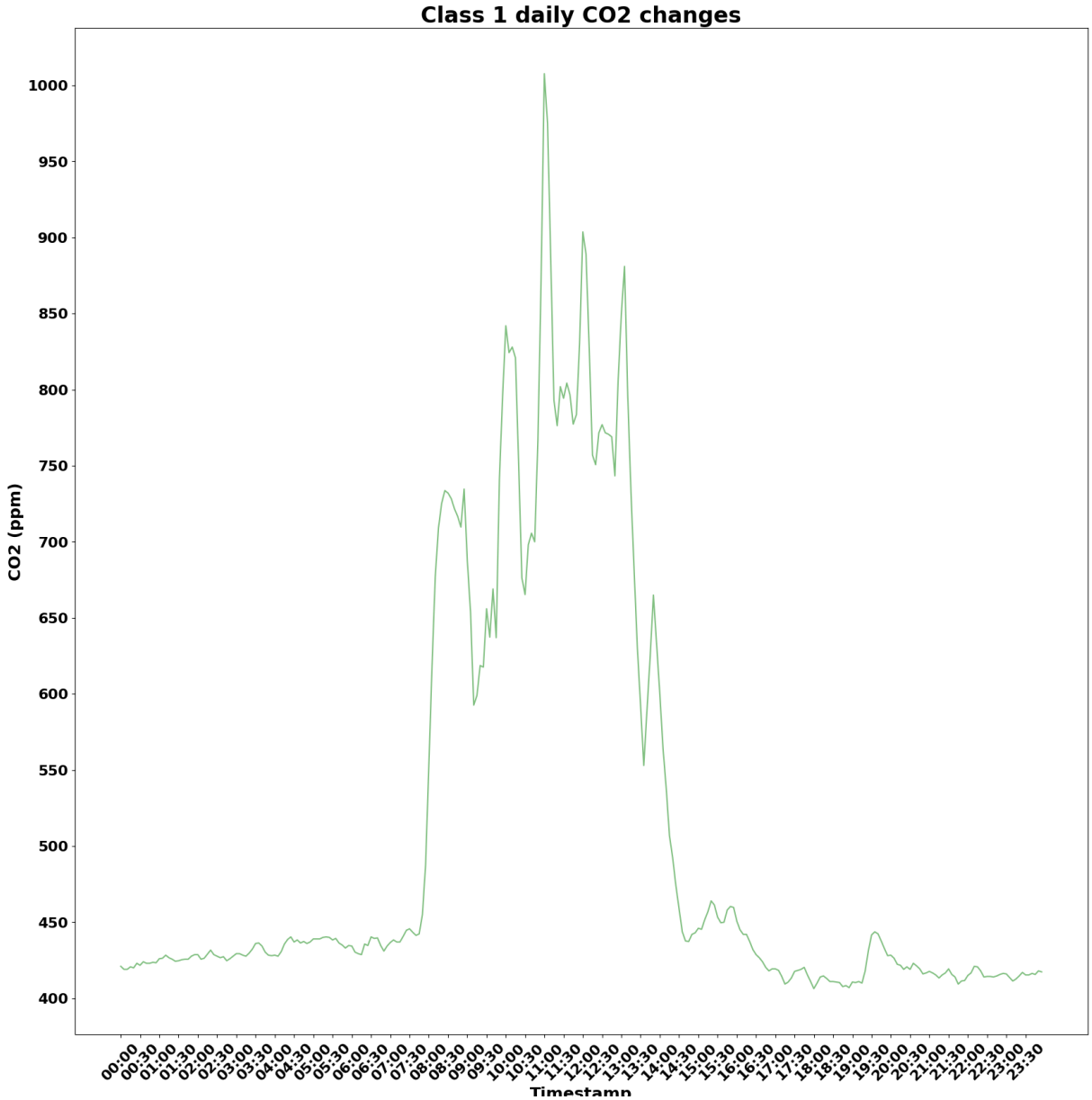


Figure 4.6 Daily plot for CO_2 fluctuations

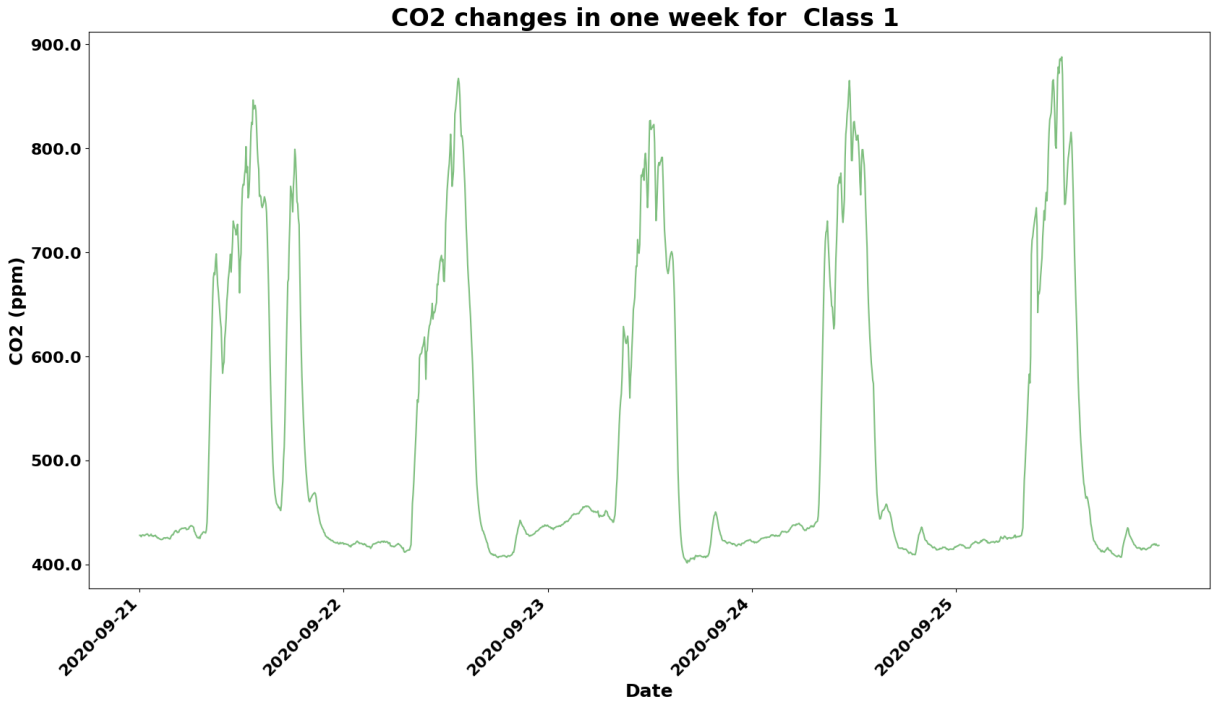


Figure 4.7 Weekly plot for CO_2 fluctuations

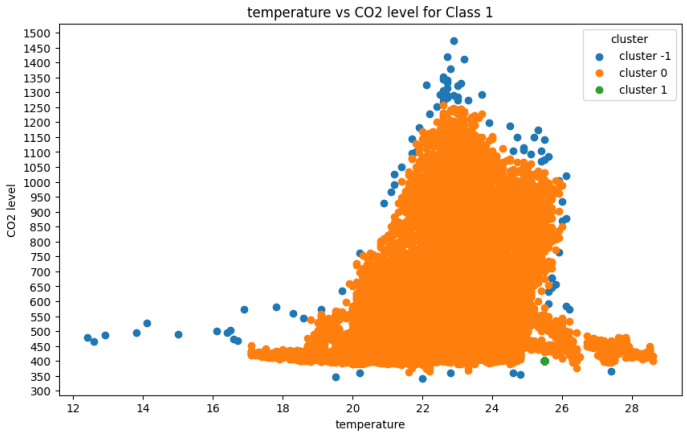
4.3.1 DBSCAN method

This method is explained in Section 2.9.2. For this method, two parameters are required which are the epsilon and minimum number of samples. These parameters are:

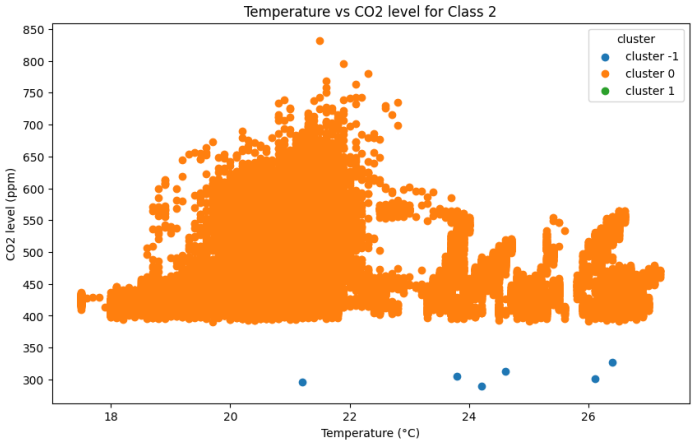
- Epsilon (eps) which denotes the maximum distance between any two data points in a cluster. If there is a distance between two points that is less than or equal to epsilon, the algorithm considers them to be in the same cluster. The degree of granularity at which the clusters are produced depends on the value of epsilon. Smaller, more compact clusters will arise from a low value of epsilon, whereas bigger, more dispersed clusters will result from a high number. This was decided upon as 0.1 based on the number of required clusters which is 3.
- Minimum samples which refer to the bare minimum number of data points necessary to create a dense zone. The minimum number of neighboring points that must exist for a point to be a core point is determined by this value and is referred to as "min_samples" in DBSCAN. For this work, a min_sample of 10 was used.

This method however performed poorly for the dataset. Figure 4.8 shows that the algorithm clustered one data point with majority of the data. Several reasons could account for this behavior. Reasons such as uneven cluster density, insufficient data, Noise in the date, etc.

(a)



(b)



(c)

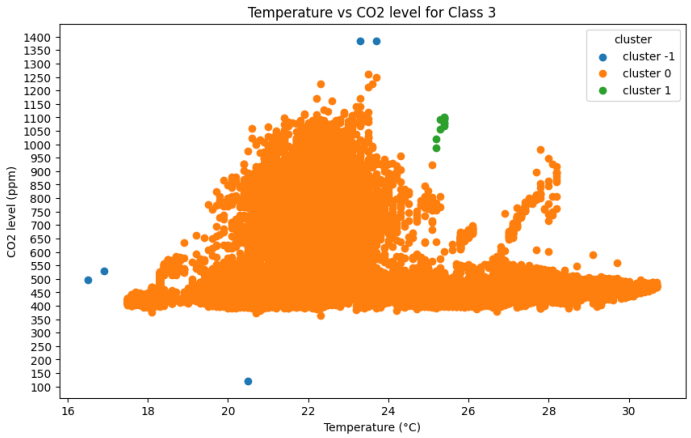
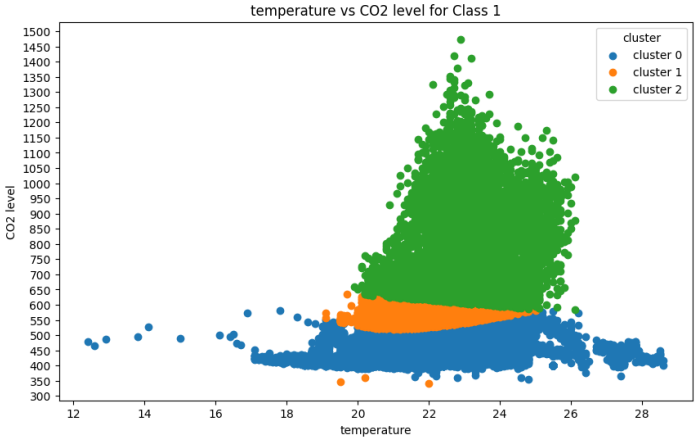


Figure 4.8 DBSCAN cluster for three classes. (a) Class 1, (b) Class 2 and (c) Class 3.

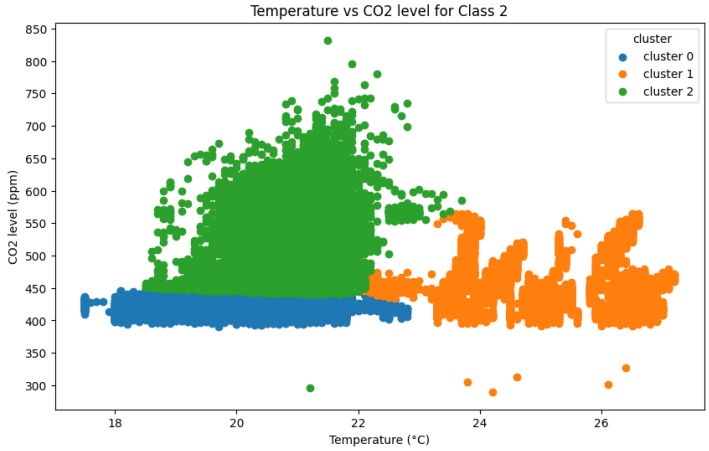
4.3.2 GMM method

The objective of GMM is to determine the underlying clusters in the data by estimating the Gaussian distributions' parameters. Each data point in a GMM is considered to belong to one of many Gaussian distributions, and the probability density function of the mixture model determines which Gaussian distribution each point belongs to. In comparison to other clustering algorithms like K-means or DBSCAN, GMM has the advantage of being able to model more complicated cluster structures and find clusters with different densities. GMM does, however, have certain drawbacks as well which includes its inability to perform well for high-dimensional data and the fact that it is computationally more expensive than k-means. Figure 4.9 shows that this method also performed quite poorly for the dataset.

(a)



(b)



(c)

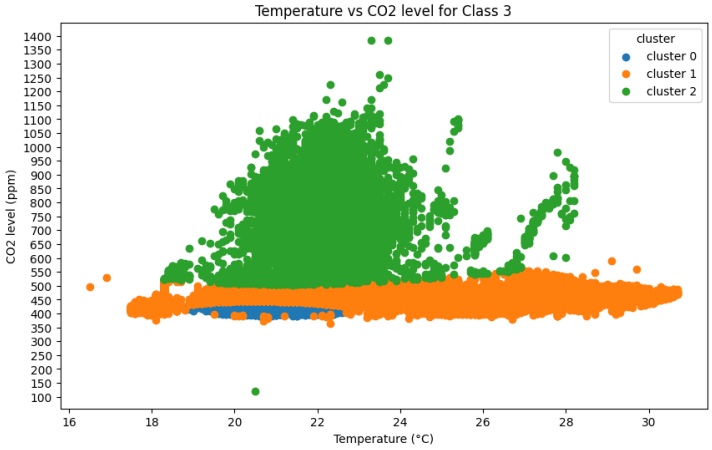


Figure 4.9 GMM cluster for 3 classes. (a) Class 1, (b) Class 2 and (c) Class 3

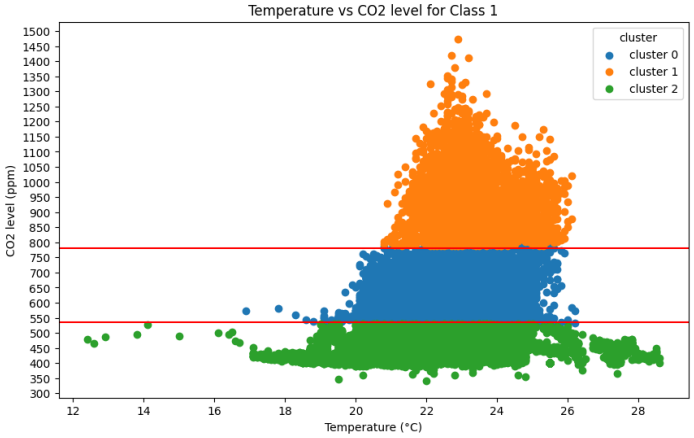
4.3.3 K-means method

This method is explained in Section 2.9.1. The only parameter required by this algorithm is the number of clusters which in the case of this study is 3. Figure 4.10 shows that this method performed best in clustering all the dataset and analyzing the clusters revealed the below results for the classes:

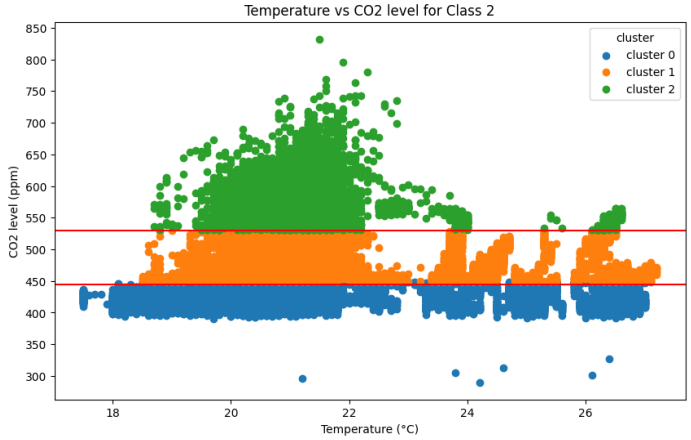
- As depicted in Figure 4.10 (a), CO_2 levels below 535ppm were clustered as low occupancy, CO_2 levels between 535ppm and 780ppm were clustered as medium occupancy and CO_2 levels above 780ppm were clustered as full occupancy for class 1.
- As depicted in Figure 4.10 (b), CO_2 levels below 445ppm were clustered as low occupancy, CO_2 levels between 445ppm and 535ppm were clustered as medium occupancy and CO_2 levels above 535ppm were clustered as full occupancy for Class 2.
- As depicted in Figure 4.10 (b), CO_2 levels below 550ppm were clustered as low occupancy, CO_2 levels between 550ppm and 785ppm were clustered as medium occupancy and CO_2 levels above 785ppm were clustered as full occupancy for Class 3.

These clusters exhibit varying levels for the three classes due to the differences in their respective areas, and hence volumes. As the area of a space increases, the diffusion of CO_2 gases into the surrounding air improves, resulting in lower measured CO_2 levels by the sensor. Consequently, a smaller class with fewer occupants may have a higher CO_2 level compared to a larger class with more occupants. However, it is important to note that higher CO_2 levels in a smaller class do not necessarily indicate higher occupancy. The CO_2 level in a space is influenced by two factors: the volume of the space and the ventilation flow rate. In the subsequent sections, both of these parameters will be utilized to validate the accuracy of the occupancy estimation method employed in this study.

(a)



(b)



(c)

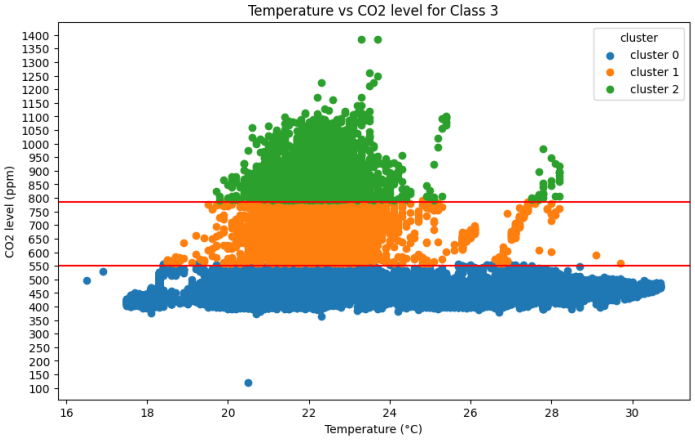


Figure 4.10 K-means cluster for 3 classes. (a) Class 1, (b) Class 2, (c) Class 3

4.4 Validation of the methodology using CO_2 mass balance equation for mixing in gaseous spaces

This method has been introduced in Section 2.6. The authors of [46] describes the dynamic method for detecting the actual occupancy in indoor spaces by measuring the indoor and outdoor CO_2 concentration and flow rate of ventilated spaces. The authors describe the parameters that affect ventilated spaces as seen in Figure 4.11 where v is the volumetric flow rate (m^3/s) at which the air enters into the space from the ventilation units, V_s is the volume (m^3) of supply air in the space, C_s is the CO_2 concentration (ppm) in supply air (if any), N is the number of occupants in the space, C_{rtn} is the CO_2 concentration (ppm) of the return air or outside air and C_R is the CO_2 concentration (ppm) in the space.

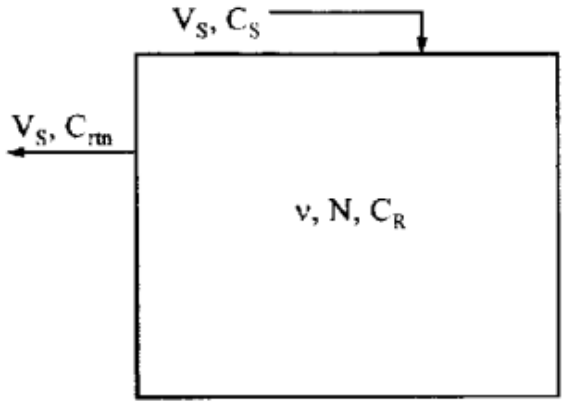


Figure 4.11 Parameters for a ventilated space [46]

The authors in [46] describe a CO_2-based occupancy detection method which aims to estimate the number of people in a room based on the measured CO_2 concentration. This method is a refined version of the traditional mass balance method, which is based on the principle that the CO_2 concentration in a room is directly proportional to the number of people in the room. It does this by using a dynamic model of the indoor air quality in the room, which considers the sources and sinks of CO_2 in the room, as well as the transport of CO_2 in the room due to air currents and mixing. The equations described by this author are however very complex and require a lot of unavailable parameters.

This problem was eliminated by the study in [7] where an equation was developed to exactly calculate the number of occupants in a space given the CO_2 level and other parameters. This equation is listed in eq (2.2) of this work. The following assumptions are valid for the equation:

- The CO_2 removal rate from the space was not taken into consideration in the model.
- The volume of air in the space is constant.
- The concentration of CO_2 in the air is uniform and well mixed within the space over time
- There is no significant source or sink of CO_2 within the space, other than the occupants' respiration and the ventilation system's intake and exhaust
- The ventilation rate is constant and uniform throughout the space, with no short-circuiting or dead zones
- The occupancy level is constant over the time period being analyzed
- There is no significant air exchange between the space and adjacent spaces or the outdoor environment

From eq. (2.2), the number of occupants N can be calculated as shown in eq. (4.1)

$$N = \frac{Q\,(C_t - C_0)}{c\left(1 - e^{\left(-\frac{Qt}{V}\right)}\right)} \tag{4.1}$$

For this study, *c*, the CO_2 generation rate per person was chosen as 0.0055L/s (39.718,8mg/h) which is an average of the values found in [47] for classrooms since student age in the building varies. The following sub-sections will illustrate how variations in the power level and flow rate of the ventilation unit can impact the number of occupants for the same CO_2 level, demonstrating that a full occupancy for one ventilation level does not equate to a full occupancy for another ventilation level. Hence, the feasibility of using electricity consumption level data to cluster ventilation levels is demonstrated. For all classes, we use a sample curve in Figure 4.12 for validation. The assumption is that every student is already in the class when classes start.

From the ventilation units' datasheet in appendix 3, the Specific Fan Power (SFP) is of importance for this calculation as it is a measure of the energy efficiency of the ventilation system in question. It is defined as the power required by the ventilation fan to move a unit of air mass through the system. The units have 2 fans, the exhaust fan and the inlet fan, and these will be taken into consideration during the calculations. The following equations will be used:

$$P_{Max,Fan} = SFP_{fan} * Q_{Max,Fan} \quad (4.2)$$

$$P_{AHU} = P_{fan} * 2 \quad (4.3)$$

$$P = R_{fan} * Q^3 \quad (4.4)$$

Where P - power of the unit under investigation, kW,

Q - flow rate of the unit under investigation, m^3s^{-1},

$P_{Max,Fan}$ - maximum fan power for the ventilation, kW,

$Q_{Max,Fan}$ - maximum fan flow rate for the ventilation, m^3s^{-1},

P_{AHU} is the power of the AHU, kW,

R_{fan} is the flow resistance.

Flow resistance is a measure of the force required to move a fluid through a pipe or duct.

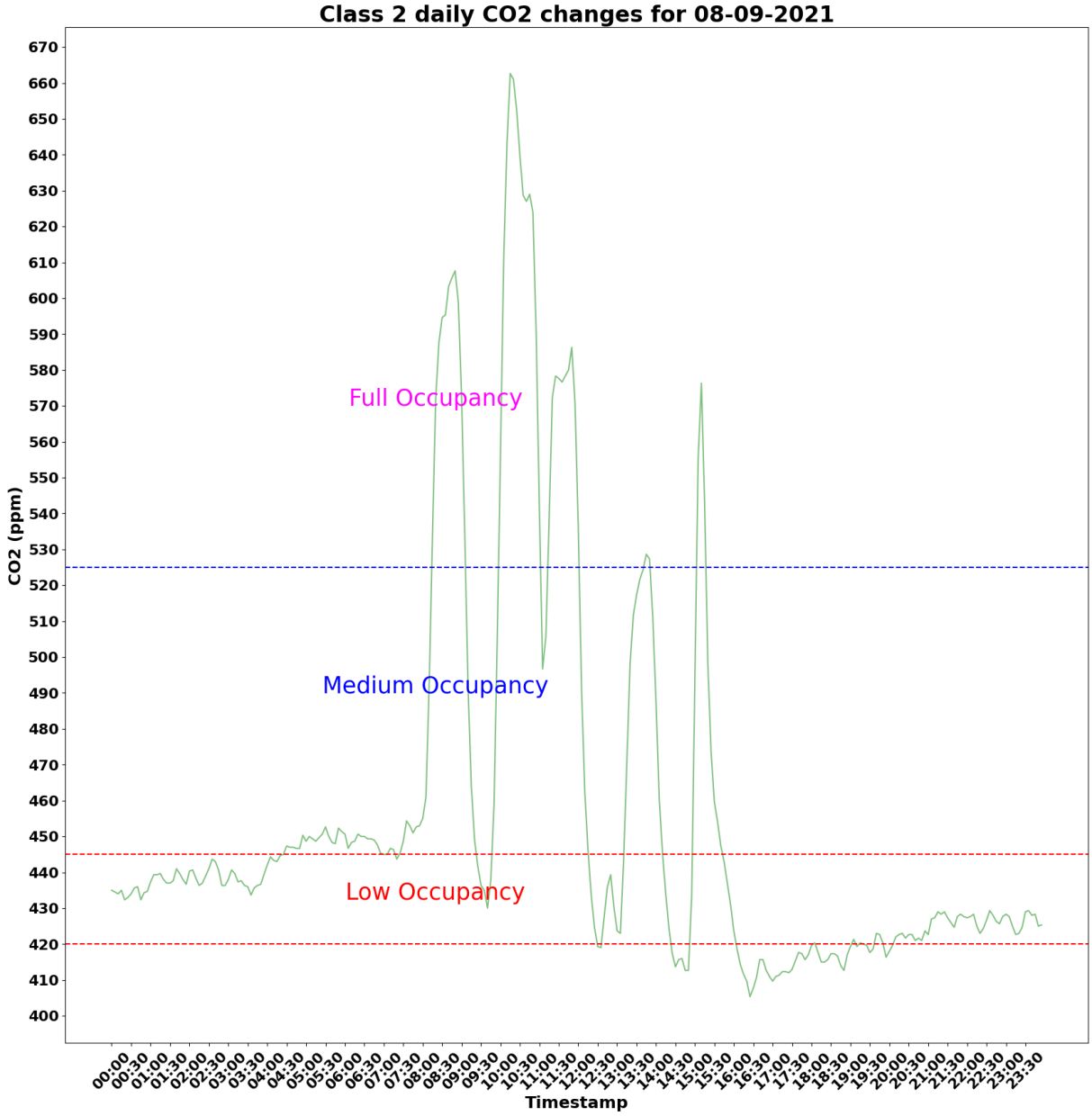


Figure 4.12 Sample plot for validation

4.4.1 Low occupancy

For the plot in Figure 4.12, the red marker represents low occupancy interval during which time 12:30 to 13:12 was selected along the x-axis and CO_2 level between 420 ppm and 445 ppm on the y-axis.

For class 1. This class uses ventilation unit SV06. From Appendix 3, for SV06 SFP_{fan} = 1,25 kW/(m^3s^{-1}) and Maximum flowrate, Q_{max} = 2,46 m^3s^{-1}.
Analyzing Figure 4.12 for CO_2 values for low occupancy on the y-axis gives values:
CO_2 level at 13:12 from dataset C_t = 445 ppm (509,8 mgm^{-3}), CO_2 level at 12:30 from dataset C_0 = 420 ppm (481,2 mgm^{-3}), Energy consumed in 24 hrs from dataset, E = 38,92 kWh, Time t = 42 mins (0,7 h), Volume of room = 153,2 m^3, CO_2 generation rate per person c = 0,0055 L s^{-1} (39.718,8 mgh^{-1}).
From eq. (4.2),

$$P_{AHU,Max,Fan} = 1{,}25 \ * \ 2{,}46 = 3 \ kW$$

From eq. (4.3),

$$P_{max,fan} \ = 3.000 \ \div 2 \ = 1{,}5 \ kW$$

From eq. (4.4),

$$R_{fan} = \ \frac{1{,}5}{2{,}46^3} = \ 0{,}1 \ kW/(m^3s^{-1})^3$$

Power consumed for this level,

$$P = \ \frac{Energy}{Time} = \ \frac{38{,}92}{24} = 1{,}62 \ kW$$

From eq. (4.3), P_{AHU} for this level

$$P_{fan} = 1{,}62 \ \div 2 = 0{,}81 \ kW$$

Flow rate for this level, Q from eq (4.4),

$$Q = \ \sqrt[3]{\frac{0{,}81}{0{,}1}} = 2 \ m^3s^{-1}$$

Q = 2 m^3s^{-1} = 7.200 m^3h^{-1}
Populating eq. (4.1) results in the below

$$N = \frac{7.200(509{,}8 - 481{,}2)}{39.718{,}8 \ \left(1 - e^{\left(-\frac{7.200 \ * \ 0{,}7}{153{,}2}\right)}\right)}$$

Solving for N gives N ≈ 6 occupants.

For class 2. This class uses ventilation unit SV05. From Appendix 3, for SV05 SFP_{fan} = 1,25 kW/(m^3s^{-1}) and Maximum flowrate, Q_{max} = 2,12 m^3s^{-1}.

Analyzing Figure 4.12 for CO_2 values for low occupancy on the y-axis gives values:

CO_2 level at 13:12 from dataset C_t = 445 ppm (509,8 mgm^{-3}), CO_2 level at 12:30 from dataset C_0 = 420 ppm (481,2 mgm^{-3}), Energy consumed in 24 hrs from dataset E = 23,55 kWh, Time t = 42 mins (0,7 h), Volume of room = 168,2 m^3, CO_2 generation rate per person c = 0,0055 L s^{-1} (39.718,8 mgh^{-1}).

From eq. (4.2),

$$P_{AHU,Max,Fan} = 1{,}25 \ * \ 2{,}12 = 2{,}65 \ kW$$

From eq. (4.3),

$$P_{max,fan} \ = 2{,}65 \div 2 \ = 1{,}3 \ kW$$

From eq. (4.4),

$$R_{fan} = \frac{1{,}3}{2{,}12^3} = \ 0{,}14 \ kW/(m^3 s^{-1})^3$$

Power consumed for this level,

$$P = \frac{Energy}{Time} = \frac{23{,}55}{24} = 0{,}98 \ kW$$

From eq. (4.3), P_{AHU} for this level

$$P_{fan} = 0{,}98 \ \div 2 = 0{,}49 \ kW$$

Flow rate for this level, Q from eq (4.4),

$$Q = \sqrt[3]{\frac{0{,}49}{0{,}14}} = 1{,}52 \ m^3 s^{-1}$$

Q = 1,52 m^3s^{-1} = 5.466 m^3h^{-1}

Populating eq. (4.1) results in the below

$$N = \frac{5.466(509{,}8 - 481{,}2)}{39.718{,}8 \left(1 - e^{\left(-\frac{5.466 * 0{,}7}{168{,}2}\right)}\right)}$$

Solving for N gives N ≈ 5 occupants.

For class 3. This class uses ventilation unit SV07. From Appendix 3, for SV05 SFP_{fan} = 1,43 kW/(m^3s^{-1}) and Maximum flowrate, Q_{max} = 3,59 m^3s^{-1}.

Analyzing Figure 4.12 for CO_2 values for low occupancy on the y-axis gives values:

CO_2 level at 13:12 from dataset C_t = 445 ppm (509,8 mgm^{-3}), CO_2 level at 12:30 from dataset C_0 = 420 ppm (481,2 mgm^{-3}), Energy consumed in 24 hrs from dataset E = 108,03 kWh, Time t = 42 mins (0,7 h), Volume of room = 200,48 m^3, CO_2 generation rate per person c = 0,0055 L/s (39.718,8 mgh^{-1}).

From eq. (4.2),

$$P_{AHU,Max,Fan} = 1{,}43 \ * \ 3{,}59 = 5{,}13 \ kW$$

From eq. (4.3),

$$P_{max,fan} \ = 5{,}13 \div 2 \ = 2{,}57 \ kW$$

From eq. (4.4),

$$R_{fan} = \frac{2{,}57}{3{,}59^3} = \ 0{,}055 \ kW/(m^3 s^{-1})^3$$

Power consumed for this level,

$$P = \frac{Energy}{Time} = \frac{49}{24} = 2{,}0 \ kW$$

From eq. (4.3), P_{AHU} for this level

$$P_{fan} = 2{,}0 \ \div 2 = 1{,}00 \ kW$$

Flow rate for this level, Q from eq (4.4),

$$Q = \sqrt[3]{\frac{1{,}00}{0{,}055}} = 2{,}60 \ m^3 s^{-1}$$

Q = 2,60 m^3s^{-1} = 9.466 m^3h^{-1}

Populating eq. (4.1) results in the below

$$N = \frac{9.466(509{,}8 - 481{,}2)}{39.718{,}8 \left(1 - e^{\left(-\frac{9.466 * 0{,}7}{200{,}48}\right)}\right)}$$

Solving for N gives N ≈ 8 occupants.

4.4.2 Medium occupancy

For the plot in Figure 4.12, the blue marker represents medium occupancy interval during which time 12:30 to 13:30 was selected along the x-axis and CO_2 level between 420 ppm and 520 ppm on the y-axis.

For class 1. Analyzing Figure 4.12 for CO_2 values on the y-axis for medium occupancy gives values:
CO_2 level at 12:30 from dataset C_t = 420 ppm (481,20 mgm^{-3}), CO_2 level at 13:30 from dataset C_0 = 520 ppm (595,70 mgm^{-3}), Time t = 1 h, from Section 4.4.1 flow rate for class 1, Q = 7.200 m^3h^{-1}
Populating eq. (4.1) results in the below

$$N = \frac{7.200\ \ (595{,}70 - 481{,}20\ )}{39.718{,}8\ \left(1 - e^{\left(-\frac{7.200 * 1}{153{,}2}\right)}\right)}$$

Solving for N gives N ≈ 21 occupants.

For class 2. Analyzing Figure 4.12 for CO_2 values on the y-axis for medium occupancy gives values:
CO_2 level at 12:30 from dataset C_t = 420 ppm (481,20 mgm^{-3}), CO_2 level at 13:30 from dataset C_0 = 520 ppm (595,70 mgm^{-3}), Time t = 1 h, from Section 4.4.1 flow rate for class 2, Q = 5.466 m^3h^{-1}
Populating eq. (4.1) results in the below

$$N = \frac{5.466\ (595{,}70 - 481{,}20\ )}{39.718{,}8\ \left(1 - e^{\left(-\frac{5.466\ * 1}{168{,}2}\right)}\right)}$$

Solving for N gives N ≈ 15 occupants.

For class 3. Analyzing Figure 4.12 for CO_2 values on the y-axis for medium occupancy gives values:
CO_2 level at 12:30 from dataset C_t = 420 ppm (481,20 mgm^{-3}), CO_2 level at 13:30 from dataset C_0 = 520 ppm (595,70 mgm^{-3}), Time t = 1 h, from Section 4.4.1 flow rate for class 3, Q = 9.466 m^3h^{-1}

Populating eq. (4.1) results in the below

$$N = \frac{9.466(595{,}70 - 481{,}20\ )}{39.718{,}8\ \left(1 - e^{\left(-\frac{9.466 * 1}{200{,}48}\right)}\right)}$$

Solving for N gives N ≈ 27 occupants.

4.4.3 High occupancy

For the plot in Figure 4.12, the magenta marker represents high occupancy interval during which time 12:30 to 13:40 was selected along the x-axis and CO_2 level between 420 ppm and 530 ppm on the y-axis.

For class 1. Analyzing Figure 4.12 for CO_2 values on the y-axis for high occupancy gives values:

CO_2 level at 12:30 from dataset C_t = 420 ppm (481,20 mgm^{-3}), CO_2 level at 13:40 from dataset C_0 = 530 ppm (607.2 mgm^{-3}), Time t = 70 mins (1,2 h), from Section 4.4.1 flow rate for class 1, Q = 7.200 m^3h^{-1}

Populating eq. (4.1) results in the below

$$N = \frac{7.200(607.2 - \ 481{,}20)}{39.718{,}8\ \left(1 - e^{\left(-\frac{7.200 * 1{,}2}{153{,}2}\right)}\right)}$$

Solving for N gives N ≈ 23 occupants.

For class 2. Analyzing Figure 4.12 for CO_2 values on the y-axis for high occupancy gives values:

CO_2 level at 12:30 from dataset C_t = 420 ppm (481,20 mgm^{-3}), CO_2 level at 13:40 from dataset C_0 = 530 ppm (607.2 mgm^{-3}), Time t = 70 mins (1,2 h), from Section 4.4.1 flow rate for class 2, Q = 5.466 m^3h^{-1}

Populating eq. (4.1) results in the below

$$N = \frac{5.466\ (607.2 - \ 481{,}20)}{39.718{,}8\ \left(1 - e^{\left(-\frac{5.466\ * 1{,}2}{168{,}2}\right)}\right)}$$

Solving for N gives N ≈ 18 occupants.

For class 3. Analyzing Figure 4.12 for CO_2 values on the y-axis for high occupancy gives values:
CO_2 level at 12:30 from dataset C_t = 420 ppm (481,20 mgm^{-3}), CO_2 level at 13:40 from dataset C_0 = 530 ppm (607.2 mgm^{-3}), Time t = 70 mins (1,2 h), from Section 4.4.1 flow rate for class 3, Q = 9.466 m^3h^{-1}
Populating eq. (4.1) results in the below

$$N = \frac{9.466(607.2 - 481{,}20)}{39.718{,}8\left(1 - e^{\left(-\frac{9.466 * 1{,}2}{200{,}48}\right)}\right)}$$

Solving for N gives N ≈ 30 occupants.

The findings of this study suggest that the proposed method of estimating occupancy by clustering environmental sensor data using historical energy consumption data of the ventilation unit and the DBSCAN method is effective. The observed variations in the calculated number of occupants irrespective of the same CO_2 level provide strong evidence to support the reliability and accuracy of the proposed method. Table 4.1 summarizes the values calculated for the number of occupants using the eq. (4.1). The occupancy for each class is given as low, medium, and high from top to bottom.

Table 4.1 Summary of parameters and occupancy level for each class

Class	Unit	Flow rate, Q, m^3h^{-1}	Vol of class, m^3	Power, P, kW	Occupancy
1	SV06	7.2	153,2	0,8	6
					21
					23
2	SV05	5.5	168,2	0,5	5
					15
					18
3	SV07	9.5	200,5	1,0	8
					27
					30

5. CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORKS

In this work, a review and evaluation of prior research on occupancy estimation was carried out, highlighting the strengths and weaknesses of different AI-based methods. The use of more recent methods, such as Wi-Fi, wireless sensors, and webcams, combined with AI techniques for occupancy studies was also discussed. However, previous literature lacked the comparison of different combinations of sensor data and their impact on occupancy estimation reduction. This study fills this gap by investigating the use of historical energy consumption data and environmental sensor data to estimate occupancy levels. Finally, the model was validated using the CO_2 mass balance equation. The findings of this study have implications for future research and practical applications in building energy management systems.

Based on the findings of this study, the following recommendations for future research on occupancy estimation can be made:

- One promising direction for future research would involve implementing the proposed solution for controlling amenities in classrooms. The identified clusters can serve as a basis for determining periods of low or no occupancy in the classes. This information can be integrated into smart dashboards or systems to notify building managers and trigger specific actions based on the occupancy levels. For instance, actions such as turning off lights and switches in unoccupied classrooms can be automated, leading to significant energy conservation and cost reduction.
- Furthermore, this solution holds potential for future applications in (DCV) systems, leveraging the cluster analysis and CO_2 levels. By dynamically adjusting ventilation rates based on the identified clusters and CO_2 levels, it becomes possible to maintain optimal IAQ while simultaneously reducing energy consumption and associated expenses. For example, according to EU regulations, indoor CO_2 levels should not exceed 1000 ppm. However, in class 1, there were recorded CO_2 values above 1200 ppm, indicating that the ventilation unit was not functioning properly as required. The utilization of cluster analysis in this study can be instrumental in identifying such deviations and facilitating necessary adjustments to ensure optimal IAQ.

6. JÄRELDUS JA SOOVITUSED TULEVASTEKS TÖÖDEKS

Käesolevas töös viidi läbi ülevaade ja hindamine varasematest uurimustest seoses hõivatuse hindamisega, rõhutades erinevate tehisintellekti meetodite tugevusi ja nõrkusi. Arutati ka viimase aja meetodite, nagu Wi-Fi, traadita andurid ja veebikaamerad, kasutamist koos tehisintellekti tehnikatega hõivatuse uuringutes. Siiski puudus varasemas kirjanduses erinevate andurite andmekombinatsioonide võrdlus ja nende mõju hõivatuse hindamisele. Käesolev uuring täidab selle lünga, uurides ajaloolise energiatarbimise andmete ja keskkonnaandurite andmete kasutamist hõivatuse taseme hindamiseks. Lõpuks valideeriti mudel CO_2 massibalansi võrrandiga. Käesoleva uuringu tulemused omavad olulisi tagajärgi tulevastele uurimustele ja praktilistele rakendustele hoonete energiavalitsemise süsteemides.

Käesoleva uuringu tulemuste põhjal saab teha järgmised soovitused tulevastele hõivatuse hindamise uurimustele:

- Tulevaste uuringute jaoks pakub üks paljulubav suund lahenduse rakendamine klassiruumide haldamiseks. Määratud klastrid saavad olla aluseks perioodide tuvastamiseks, mil klassides on madal või puudub hõivatus. See teave saab integreerida nutikatesse juhtpaneelidesse või süsteemidesse, et teavitada hoonehaldureid ja käivitada konkreetseid tegevusi sõltuvalt hõivatuse tasemest. Näiteks saab automatiseerida valguste ja lülitite väljalülitamise tühjades klassiruumides, mis toob kaasa olulise energiasäästu ja kulude vähenemise.
- Lisaks sellele omab see lahendus potentsiaali tulevasteks rakendusteks Nõudluspõhise ventilatsiooni (DCV) süsteemides, kasutades ära klastrite analüüsi ja CO_2 taset. Klastrite ja CO_2 taseme põhjal ventilatsioonimäära dünaamiline kohandamine võimaldab säilitada optimaalset siseõhu kvaliteeti (IAQ), samal ajal vähendades energiatarbimist ja sellega seotud kulusid. Näiteks vastavalt Euroopa Liidu seadustele ei tohiks siseruumides CO_2 tase ületada 1000 ppm. Siiski, klassis 1 registreeriti CO_2 väärtusi üle 1200 ppm, mis näitab, et ventilatsioonisüsteem ei töötanud nõuetekohaselt. Käesoleva uuringu raames läbiviidud klastrianalüüs võib olla oluline abivahend selliste kõrvalekallete tuvastamisel ning vajalike muudatuste tegemisel optimaalse siseõhu kvaliteedi tagamiseks. See rõhutab klastrianalüüsi kasutamise olulisust siseõhu kvaliteedi juhtimisel ja vastavuse tagamisel regulatiivsetele standarditele.

7. REFERENCES

[1] C. Kanthila, A. Boodi, K. Beddiar, Y. Amirat, and M. Benbouzid, "Building Occupancy Behavior and Prediction Methods: A Critical Review and Challenging Locks," *IEEE Access*, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 79353–79372, 2021. doi: 10.1109/ACCESS.2021.3083534.

[2] I. B. Arief-Ang, F. D. Salim, and M. Hamilton, "CD-HOC: Indoor Human Occupancy Counting using Carbon Dioxide Sensor Data," 2017. doi: https://doi.org/10.48550/arXiv.1706.05286.

[3] P. Liu, M. Justo Alonso, and H. M. Mathisen, "Heat recovery ventilation design limitations due to LHC for different ventilation strategies in ZEB," *Build Environ*, vol. 224, p. 109542, 2022, doi: 10.1016/j.buildenv.2022.109542.

[4] M. Amayri, A. Arora, S. Ploix, S. Bandhyopadyay, N. Quoc-Dung, and V. R. Badarla, "Estimating occupancy in heterogeneous sensor environment," *Energy Build*, vol. 129, pp. 46–58, Oct. 2016, doi: 10.1016/j.enbuild.2016.07.026.

[5] A. Franco and F. Leccese, "Measurement of CO2 concentration for occupancy estimation in educational buildings with energy efficiency purposes," *Journal of Building Engineering*, vol. 32, Nov. 2020, doi: 10.1016/j.jobe.2020.101714.

[6] Y. Longqi, K. Ting, and M. B. Srivastava, "Inferring occupancy from opportunistically available sensor data," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, IEEE, Mar. 2014, pp. 60–68. doi: 10.1109/PerCom.2014.6813945.

[7] S. Kar and P. K. Varshney, "Accurate Estimation of Gaseous Strength Using Transient Data," *IEEE Trans Instrum Meas*, vol. 60, no. 4, pp. 1197–1205, Apr. 2011, doi: 10.1109/TIM.2010.2084731.

[8] B. Dong *et al.*, "An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network," *Energy Build*, vol. 42, no. 7, pp. 1038–1046, Jul. 2010, doi: 10.1016/j.enbuild.2010.01.016.

[9] "REHVA EPB Standard." https://www.rehva.eu/activities/epb-center-on-standardization/epb-standards-energy-performance-of-buildings-standards (accessed Nov. 15, 2022).

[10] M. S. Zuraimi, A. Pantazaras, K. A. Chaturvedi, J. J. Yang, K. W. Tham, and S. E. Lee, "Predicting occupancy counts using physical and statistical Co2-based modeling methodologies," *Build Environ*, vol. 123, pp. 517–528, Oct. 2017, doi: 10.1016/j.buildenv.2017.07.027.

[11] Guillaume Ansanay-Alex, "Estimating Occupancy Using Indoor Carbon Dioxide Concentrations Only in an Office Building: a Method and Qualitative Assessment," 2013, pp. 1–9. Accessed: Dec. 11, 2022. [Online]. Available: https://www.researchgate.net/publication/255739151_Estimating_Occupancy_Using_Indoor_Carbon_Dioxide_Concentrations_Only_in_an_Office_Building_a_Method_and_Qualitative_Assessment

[12] E. N. Issam and M. J. Murphy, "What Is Machine Learning?," in *Machine Learning in Radiation Oncology*, Cham: Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

[13] B. Mahesh, "Machine Learning Algorithms-A Review Machine Learning," *International Journal of Science and Research*, 2018, doi: 10.21275/ART20203995.

[14] C. Brennan, G. W. Taylor, and P. Spachos, "Designing learned CO2-based occupancy estimation in smart buildings," *IET Wireless Sensor Systems*, vol. 8, no. 6, pp. 249–255, Dec. 2018, doi: 10.1049/iet-wss.2018.5027.

[15] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/IJCTT-V48P126.

[16] Salim Dridi, "Unsupervised Learning - A Systematic Literature Review," 2021, Accessed: Dec. 12, 2022. [Online]. Available: https://www.researchgate.net/publication/357380639_Unsupervised_Learning_-_A_Systematic_Literature_Review

[17] A. C. Michalos, Ed., *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer Netherlands, 2014. doi: 10.1007/978-94-007-0753-5.

[18] K. Alanne, "A novel performance indicator for the assessment of the learning ability of smart buildings," *Sustain Cities Soc*, vol. 72, p. 103054, Sep. 2021, doi: 10.1016/j.scs.2021.103054.

[19] W. Wang, J. Chen, and T. Hong, "Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings," *Autom Constr*, vol. 94, pp. 233–243, Oct. 2018, doi: 10.1016/j.autcon.2018.07.007.

[20] K. C. J. Simma, A. Mammoli, and S. M. Bogus, "Real-time occupancy estimation using WiFi network to optimize HVAC operation," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 495–502. doi: 10.1016/j.procs.2019.08.069.

[21] P. O. Fanger, "Thermal Comfort Analysis and Applications in Environmental Engineering", doi: https://doi.org/10.1177/146642407209200337.

[22] Q. Chai, H. Wang, Y. Zhai, and L. Yang, "Using machine learning algorithms to predict occupants' thermal comfort in naturally ventilated residential buildings," *Energy Build*, vol. 217, p. 109937, Jun. 2020, doi: 10.1016/j.enbuild.2020.109937.

[23] W. Hu, Y. Wen, K. Guan, G. Jin, and K. J. Tseng, "iTCM: Toward Learning-Based Thermal Comfort Modeling via Pervasive Sensing for Smart Buildings," *IEEE Internet Things J*, vol. 5, no. 5, pp. 4164–4177, Oct. 2018, doi: 10.1109/JIOT.2018.2861831.

[24] V. L. Erickson, M. Á. Carreira-Perpiñán, and A. E. Cerpa, "Occupancy modeling and prediction for building energy management," in *ACM Transactions on Sensor Networks*, Association for Computing Machinery, 2014. doi: 10.1145/2594771.

[25] S. H. Ryu and H. J. Moon, "Development of an occupancy prediction model using indoor environmental data based on machine learning techniques," *Build Environ*, vol. 107, pp. 1–9, Oct. 2016, doi: 10.1016/j.buildenv.2016.06.039.

[26] M. Chammas, A. Makhoul, and J. Demerjian, "An efficient data model for energy prediction using wireless sensors," *Computers & Electrical Engineering*, vol. 76, pp. 249–257, Jun. 2019, doi: 10.1016/j.compeleceng.2019.04.002.

[27] X. Xu, W. Wang, T. Hong, and J. Chen, "Incorporating machine learning with building network analysis to predict multi-building energy use," *Energy Build*, vol. 186, pp. 80–97, Mar. 2019, doi: 10.1016/j.enbuild.2019.01.002.

[28] S. Lu, W. Wang, C. Lin, and E. C. Hameen, "Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE RP884," *Build Environ*, vol. 156, pp. 137–146, Jun. 2019, doi: 10.1016/j.buildenv.2019.03.010.

[29] P. W. Tien, S. Wei, J. Darkwa, C. Wood, and J. K. Calautit, "Machine Learning and Deep Learning Methods for Enhancing Building Energy Efficiency and Indoor Environmental Quality – A Review," *Energy and AI*, vol. 10. Elsevier B.V., Nov. 01, 2022. doi: 10.1016/j.egyai.2022.100198.

[30] Y. Ding, S. Han, Z. Tian, J. Yao, W. Chen, and Q. Zhang, "Review on occupancy detection and prediction in building simulation," *Building Simulation*, vol. 15, no. 3. Tsinghua University, pp. 333–356, Mar. 01, 2022. doi: 10.1007/s12273-021-0813-8.

[31] Y. Zhao and X. Zhou, "K-means Clustering Algorithm and Its Improvement Research," *J Phys Conf Ser*, vol. 1873, no. 1, p. 012074, Apr. 2021, doi: 10.1088/1742-6596/1873/1/012074.

[32] X. Jin and J. Han, "K-Means Clustering," in *Encyclopedia of Machine Learning and Data Mining*, Boston, MA: Springer US, 2017, pp. 695–697. doi: 10.1007/978-1-4899-7687-1_431.

[33] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, IEEE, Sep. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

[34] J. Jang and H. Jiang, "DBSCAN++: Towards fast and scalable density clustering," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.13105

[35] I. M. Khater, I. R. Nabi, and G. Hamarneh, "A Review of Super-Resolution Single-Molecule Localization Microscopy Cluster Analysis and Quantification Methods," *Patterns*, vol. 1, no. 3, p. 100038, Jun. 2020, doi: 10.1016/j.patter.2020.100038.

[36] "DBSCAN Clustering Algorithm in Machine Learning." https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html (accessed May 08, 2023).

[37] M. Abdallah and O. Elkeelany, "A Survey on Data Acquisition Systems DAQ," in *2009 International Conference on Computing, Engineering and Information*, IEEE, Apr. 2009, pp. 240–243. doi: 10.1109/ICC.2009.24.

[38] "Smart temp Australia (2020). Indoor Air Quality Monitor." https://smarttemp.com.au/download/102/installation-manuals/3170/smt-iaq3-co2-sensor-manual-ver-1-5.pdf (accessed Apr. 24, 2023).

[39] G. Y. Lee, L. Alzamil, B. Doskenov, and A. Termehchy, "A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance," Sep. 2021, doi: https://doi.org/10.48550/arXiv.2109.07127.

[40] "Cleaning Big Data: Most time-consuming, least enjoyable data science task." https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=331a70976f63. (accessed Mar. 28, 2023).

[41] H. Kang, "The prevention and handling of the missing data," *Korean J Anesthesiol*, vol. 64, no. 5, p. 402, 2013, doi: 10.4097/kjae.2013.64.5.402.

[42] D. Cousineau and S. Chartier, "Outliers detection and treatment: a review.," *Int J Psychol Res (Medellin)*, vol. 3, no. 1, pp. 58–67, Jun. 2010, doi: 10.21500/20112084.844.

[43] H. Liu, C. Chen, Y. Li, Z. Duan, and Y. Li, "Individual behavior analysis and trajectory prediction," in *Smart Metro Station Systems*, Elsevier, 2022, pp. 59–76. doi: 10.1016/B978-0-323-90588-6.00003-2.

[44] "https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/." https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/ (accessed May 03, 2023).

[45] "Elbow Method." https://www.csias.in/discuss-the-elbow-method/ (accessed May 06, 2023).

[46] Shengwei Wang and Xinqiao Jin, "CO2-based occupancy detection for on-line outdoor air flow control," *Indoor and Built Environment*, vol. 7, no. 3, pp. 165–181, May 1998, doi: 10.1177/1420326X9800700305.

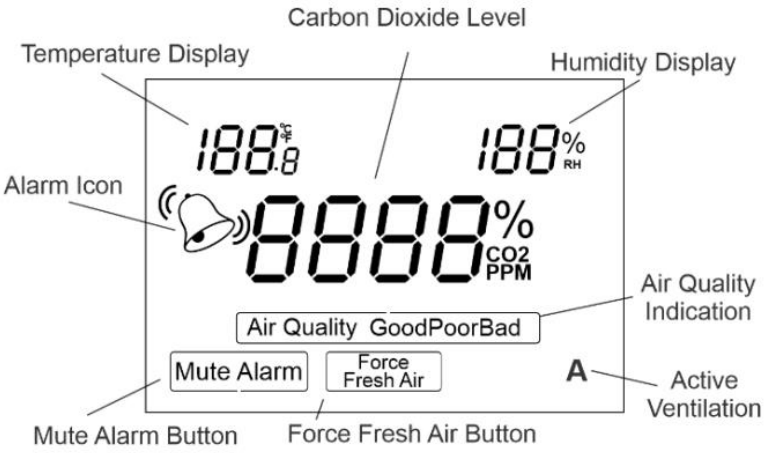
[47] A. Persily, "Development and application of an indoor carbon dioxide metric," *Indoor Air*, vol. 32, no. 7, Jul. 2022, doi: 10.1111/ina.13059.

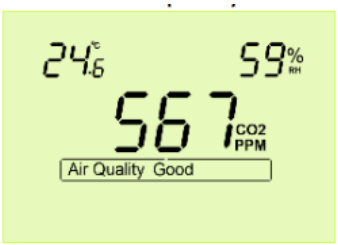
8. APPENDICES

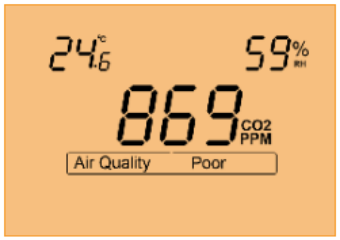
Smart Temp

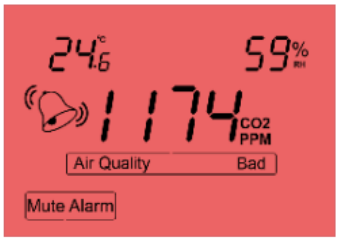
SMT-IAQ3

Indoor Air Quality Monitor

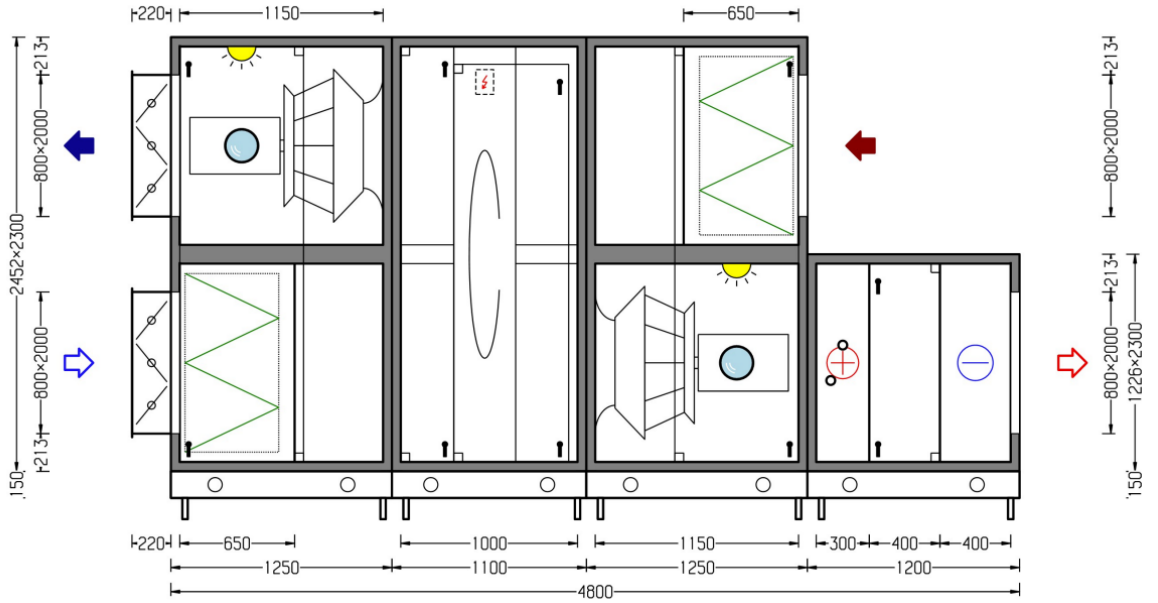


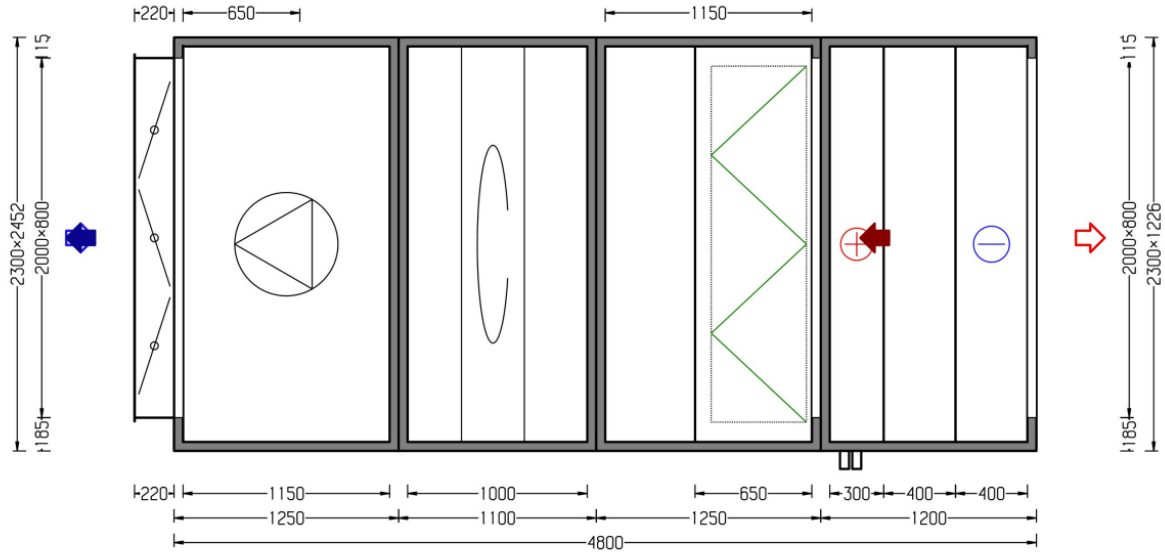






Appendix 2. Ventilation Layout





Appendix 3. Ventilation parameters

- SV05

Supply air volume flow rate	2.12 m³/sec
External static pressure	210 Pa
Mains electricity	3x400VAC±10%+N+PE, 50Hz
Specific el. power demand (SFPv)	1.25 kW/(m³/s)
Ref. density	1.2 kg/m³

- SV06

Supply air volume flow rate	2.46 m³/sec
External static pressure	220 Pa
Mains electricity	3x400VAC±10%+N+PE, 50Hz
Specific el. power demand (SFPv)	1.25 kW/(m³/s)
Ref. density	1.2 kg/m³

- SV07

Supply air volume flow rate	3.59 m³/sec
External static pressure	220 Pa
Mains electricity	3x400VAC±10%+N+PE, 50Hz
Specific el. power demand (SFPv)	1.43 kW/(m³/s)
Ref. density	1.2 kg/m³