Machine Learning and Ontology Supported Comprehensive Building Comfort Framework Design

Sisi Bie

School of Engineering

Cardiff University

A thesis submitted to Cardiff University for the degree of Doctor of Philosophy

December 2023

Abstract

This dissertation presents a comprehensive study integrating Building Information Modelling (BIM), machine learning, and ontology to enhance building comfort and energy efficiency. It explores the current state of BIM, the advancements in machine learning, and the critical dimensions of building comfort, establishing a comprehensive comfort framework that addresses the multifaceted nature of designing sustainable and comfortable buildings. The study is motivated by the need for a holistic approach to building design that efficiently balances energy consumption with occupant comfort while leveraging the potential of advanced technologies like ontology for structured knowledge representation and dynamic reasoning.

The methodology adopted in this research involves a structured approach to data generation, leveraging BIM's rich repository of building data and the predictive capabilities of machine learning. The study develops a comprehensive comfort framework through quantitative and qualitative methods, normalizing and standardizing various comfort dimensions into a unified metric that can be used to assess and compare the comfort levels of different building designs. This metric moves beyond traditional assessments, typically focused on thermal comfort, to include factors such as acoustic, visual comfort, and air quality, thereby providing a more holistic view of occupant comfort.

The dissertation further incorporates ontology to create a dynamic and adaptable comfort assessment system. By integrating Semantic Web Rule Language (SWRL) and ontology-based reasoning, the study enhances the model’s ability to evaluate real-time comfort data, predict future conditions, and suggest system optimizations. This ontology-based framework also enables the customization of comfort profiles according to user preferences, allowing for a more personalized assessment and adjustment of comfort parameters.

Additionally, the integration of machine learning with BIM and ontology is explored to revolutionize traditional building design and performance analysis. Machine learning algorithms, such as Linear Regression (LR), Artificial Neural Networks (ANN), and Random Forests (RF), are utilized to predict building performance and optimize comfort and energy efficiency strategies. This approach not only enhances the accuracy and efficiency of performance predictions but also reduces reliance on time-consuming and often biased traditional methods.

The research addresses a significant gap in the literature by proposing and validating a BIM-based machine learning and ontology engine. This engine provides a robust, dynamic, and comprehensive analysis of building data, leading to more informed decision-making in the design phase. The proposed system is tested and validated through a series of case studies, demonstrating its potential to transform building design and management practices. The dissertation concludes with a reflection on the research findings, discussing the implications for the construction industry and outlining future research directions. It emphasizes the need for continuous improvement and innovation in the integration of BIM, machine learning, and ontology, advocating for their adoption in creating more sustainable, comfortable, and efficient buildings. This study contributes to the field by offering a novel approach to building design, emphasizing the importance of holistic comfort and energy efficiency through the integration of advanced technologies.

Keywords: Comprehensive comfort framework, Machine learning, BIM model, building energy analysis, Building performance, Comfort level, Artificial neuron network, ontology.

Acknowledgments

I extend my deepest gratitude to Professor Haijiang Li, my supervisor, for his unwavering support and guidance throughout my academic journey. His expertise and insightful feedback have been invaluable, consistently guiding me back on course whenever I have strayed. His encouragement and patience have not only enhanced my academic work but also inspired me to persevere during challenging times, making a profound impact on my personal growth and academic development.

I am grateful to the dedicated staff at Cardiff University's Research Office and the Student Support services for their invaluable assistance throughout my studies. Special thanks go to Aderyn and Kristina for their understanding and support, which have been instrumental in my academic journey.

I owe a deep debt of gratitude to my mother, Hui, and my father, Ming. I thank them for the gift of life, which has given me the opportunity to experience and navigate this world.

Lastly, I extend my gratitude to myself for having the courage to walk this path and persevere through all the challenges. This journey has been my own, and I am grateful for every step that has led me here.

Sisi Bie 2023

1 Contents

List of Figures

List of Tables

1. General Introduction

1.1 Research Background

In the global effort to enhance energy efficiency and sustainability, energy consumption in the building sector has become a growing concern, with buildings contributing to approximately one-third of total energy use in China and nearly 40% in Europe (Bull, Chang, & Fleming, 2012). However, this challenge extends beyond energy conservation; it also includes the often-overlooked aspect of building comfort, which is crucial for the health and well-being of occupants. As awareness of its impact increases, comfort has emerged as a key consideration in architectural design. At the same time, decisions regarding building design are crucial in achieving optimal building performance (Eleftheriadis et al., 2017). The building design process is complex and iterative, involving the collaboration of architects, engineers, and experts from various fields. Architects, structural engineers, energy specialists, and other professionals work together to consider multiple factors at the design stage, including safety, sustainability, energy efficiency, and client requirements. These efforts also reveal a crucial aspect of building design that was often overlooked in the past—building comfort for occupants.

Building comfort includes various aspects, such as temperature, lighting, air quality, and acoustics, all of which directly affect occupants' health, productivity, and quality of life.

In contemporary society, the majority of individuals spend a significant amount of time indoors, a trend that intensified during the COVID-19 pandemic due to the increased need for remote work. In 2018, approximately 12% of employees in Germany worked from home at least occasionally, compared to over 30% in countries such as the Netherlands, Finland, Iceland, Luxembourg, and Denmark (Eurostat, 2020). This proportion experienced a notable increase during the lockdown period, though the magnitude of this growth varied among different demographic groups. Following the implementation of social distancing policies, many organizations adopted remote work arrangements to mitigate the spread of the virus. Data from April 2020 reveals that 26% of German employees worked entirely from home, while 35% engaged in a hybrid model combining both remote and on-site work （ibid）. These data emphasize the critical role of a comfortable home environment in maintaining consistent work performance and improving overall quality of life.

Furthermore, a comfortable indoor building environment has significant long-term implications for promoting both physical and mental health, as well as enhancing overall quality of life. Environmental issues such as improper temperature regulation, poor air quality, insufficient lighting, and noise pollution can adversely affect an individual's health and well-being. Therefore, incorporating comprehensive building comfort into the design phase is not merely about improving a building's energy efficiency but also about enhancing the quality of life for its occupants.

Building comfort management faces numerous challenges, such as balancing temperature, lighting, and air quality while considering individual preferences. However, machine learning plays a crucial role in addressing these issues by optimizing energy efficiency, predicting occupant needs, and enhancing system responsiveness. Building comfort encompasses thermal, visual, and acoustic factors, as well as indoor air quality. Traditional methods often struggle to balance these dynamic and sometimes conflicting factors. Machine learning models can predict occupant preferences and optimize comfort parameters in real-time. Additionally, buildings are complex, dynamic systems characterized by nonlinear relationships between weather, occupancy, and other factors. Machine learning can capture these nonlinear interactions and forecast energy-efficient strategies for maintaining comfort. Furthermore, occupants exhibit varying comfort preferences, which are challenging to model using traditional methods. Machine learning can learn individual comfort preferences and dynamically adjust environmental conditions. Achieving a balance between comfort and energy savings poses a significant challenge for buildings, but machine learning can optimise HVAC systems to minimize energy consumption while maintaining comfort. Simulating building performance is also computationally intensive and highly complex. Machine learning can generate faster, approximate models for simulations, enabling real-time optimization. However, challenges remain in applying machine learning to building comfort, such as data availability, model transparency, high implementation costs, and generalization across buildings. Overall, machine learning has the potential to significantly enhance comfort management, but it requires careful adaptation to specific building conditions. By analysing extensive and complex datasets, machine learning can help predict and optimise a building's comfort performance from the design stage, achieving dual optimisation of energy and comfort.

Thus, this research aims to develop a comprehensive building comfort design framework supported by machine learning. It accounts for the building’s energy efficiency and operational performance while adapting to the diverse and dynamic nature of building usage. In particular, in an era where people are spending more time indoors than ever, effectively enhancing the comfort of the built environment is crucial for promoting both individual and societal well-being.

In conclusion, as society advances and the nature of work and living continues to evolve, particularly in the post-pandemic era, the design of buildings must not only address traditional concerns such as aesthetics and durability but also prioritize the health and comfort of occupants. This research aims to contribute to this imperative by providing a validated, holistic approach to building design that integrates comfort, efficiency, and sustainability.

1.2 Problem Statement

Despite the potential benefits of integrating machine learning in building energy design, such as increased efficiency and optimized performance, several challenges hinder its widespread adoption. BIM is a novel way of working that has been increasingly utilised in construction industry. More and more BIM models are being developed for various organizations, and the data embedded in BIM models are massive, including geometry, materials, energy specifications and information related to health & safety. The data, information and knowledge embedded in these BIM models can be leveraged to conclude knowledge to benefit the organizations, but the detailed approach remains relatively under investigated. The complexity of building dynamics, the need for large and diverse datasets, and the integration of these advanced methods into existing design practices pose significant challenges. Additionally, there's a lack of understanding of how machine learning can not only predict but also holistically improve energy performance in varying climates, building types, and usage patterns.

Room comfort is highly related to human health, including both mental and physical well-being, particularly in residential buildings. During the pandemic and lockdown periods, almost everyone worked from home and stayed indoors. However, there have been few studies focused on residential buildings. The possible reasons are:

1) Multiple Building Styles: In the UK, residential buildings include a variety of styles such as dwellings, houses, apartments, and so on.
2) Different Room Functions: Different rooms, such as the bedroom, living room, and kitchen, have varying comfort requirements depending on their specific functions.
3) Complexity of Occupants' Schedules: Residential buildings, unlike schools and offices with relatively stable schedules, must accommodate the diverse routines of their occupants. Factors such as age, work schedules, health conditions, family size, and personal preferences make it challenging to account for comfort levels during the initial design stage.
4) Challenges in Data Collection: It is unrealistic to conduct experiments to simulate room comfort conditions. Experimental conditions are difficult to standardize, making it hard to ensure fairness and objectivity. Additionally, sensors are expensive, and collecting data over several years is time-consuming.

Given concerns about the limited scope of current comfort assessment systems, a more comprehensive framework is required to manage multiple comfort aspects during the design stage. Very few studies consider the entire process of comprehensive comfort design; therefore, it is essential to develop an integrated approach that addresses all dimensions of comfort to ensure holistic and sustainable building environments.

The technical limitations of current machine learning applications in building design include algorithmic complexity, model interpretability issues, and data security and privacy concerns. These challenges necessitate a deeper dive into the technicalities of machine learning to ensure that the models are both accurate and understandable by the end users, who are often architects and engineers with varied levels of technical expertise. Meanwhile, the social and economic factors influencing the adoption of machine learning in building design cannot be overlooked. High initial costs, a lack of skilled professionals, and the need for substantial training are all barriers that need addressing. Understanding and mitigating these factors is crucial for enabling the widespread adoption of machine learning in building energy design, ultimately leading to more sustainable and comfortable living environments. A comprehensive exploration of machine learning's role in building design, paving the way for innovative solutions that enhance building performance while considering the comfort and well-being of its occupants.

1.3 Research Objective

The primary objective of this research is to address the challenges of managing and improving comfort and energy efficiency in residential buildings, considering the dynamic, uncertain, and complex nature of room functions and occupancy preferences. The aim is to develop a comprehensive framework that not only incorporates various constraints at each stage of optimization but also utilizes high computational power to deliver near real-time results for maximum comfort and energy efficiency.

The focus will be on creating a comprehensive comfort design approach that caters to both occupants and designers. This approach should be capable of solving complex multi-objective problems within the indoor comfort domain. Recent advancements in data science and machine learning techniques demonstrate significant potential to leverage diverse data sources for predicting energy consumption and comfort levels in each room of a building. This research aims to improve computational speed and reduce redundancy by extract data from BIM models of existing buildings for the machine learning (ML) engine.

A major aim is to develop a machine learning-supported comprehensive comfort framework for building design. This framework will strive to optimize energy efficiency and operational performance while accommodating the diverse and dynamic nature of building usage. The specific goals include developing predictive models for building comfort during the design stage, and providing immediate feedback to architects and engineers for modifying the design. Furthermore, the research will explore integrating different comfort indicators, such as thermal comfort, lighting, air quality, and acoustics, into a unified design framework to ensure a multidimensional improvement in occupant comfort and building energy efficiency.

By achieving these objectives, the research aims to provide a practical and adaptable framework that not only addresses and resolves current design challenges but also anticipates and adapts to future developments. The ultimate goal is to bring about a revolutionary change in the building design industry, leading to more sustainable and human-centric living environments. Through this research, we aspire to make significant contributions to the field by offering a validated, comprehensive approach to building design that integrates comfort, efficiency, and sustainability in a harmonious manner.

1.4 Research Hypothesis

The hypothesis underpinning this research is that the application of machine learning techniques can significantly enhance the accuracy of energy consumption predictions and comfort levels during the building design stage by leveraging data from BIM models, surpassing the capabilities of traditional simulation methods. We propose that integrating data-driven insights derived from machine learning into the building design process can lead to substantial improvements in both design efficiency and building performance. Machine learning, with its ability to analyse and learn from large, diverse datasets, can detect patterns and make predictions that traditional analysis methods may not easily identify.

It is anticipated that machine learning algorithms will optimize building design by accurately predicting comfort levels, thereby enabling designers, architects, and engineers to engage in more effective and time-efficient collaboration during the design phase. Moreover, the hypothesis extends to the notion that a comprehensive comfort design framework supported by machine learning will not only address energy efficiency but also enhance overall occupant comfort. By incorporating various aspects of comfort—such as thermal, acoustic, visual comfort, and air quality—this framework aims to create buildings that are not only energy-efficient but also conducive to health and well-being.

The validation of this hypothesis will involve rigorous testing and analysis of machine learning models in diverse building scenarios, benchmarking their performance against traditional simulation methods. If proven successful, this hypothesis could drive a transformative shift in building design practices, encouraging the broader adoption of machine learning techniques in the industry and contributing to the creation of more efficient, comfortable, and sustainable living and working environments

1.5 Overview of Methodology

This research adopts a systematic approach to developing a comprehensive framework aimed at enhancing both building comfort and energy efficiency. The methodology involves a thorough investigation of relevant standards and literature, along with the collection of extensive data from BIM models at the design stage of buildings. This data encompasses various detailed types, including geometrical, material, operational, and environmental information, all of which undergo rigorous preprocessing to ensure consistency and usability for subsequent analysis.

Appropriate machine learning algorithms are selected based on specific requirements for optimizing building comfort and energy efficiency. The process includes the tuning and validation of models to ensure accuracy and reliability. A comparative analysis of different input parameters and algorithms will identify the most effective combinations for predicting and enhancing building performance. To address the comprehensive nature of the comfort framework, the research adopts a standardization normalization method. It involves the study of different comfort aspects – thermal, acoustic, visual comfort, and air quality – culminating in the calculation of a unified total building comfort index. This index allows for a holistic assessment of comfort levels, integrating various comfort parameters into a quantifiable standard. The research delineates detailed evaluation criteria based on international and domestic regulatory standards, providing quantifiable benchmarks for assessing the comfort level of buildings. These criteria serve as a guideline for designers and engineers to determine whether their designs meet the expected comfort requirements. The methodology calculates the comfort levels for different areas or rooms and weights them according to their area ratios to derive a total comfort quantification for the entire building. This weighted approach allows for a detailed assessment of individual areas as well as an overall comfort evaluation for the entire structure. Designers and engineers can use this comprehensive assessment to optimize designs, ensuring that the building's comfort level meets or exceeds regulatory standards.

The validation of the methodology involves rigorous testing and analysis of machine learning models in diverse building design scenarios, comparing their performance with traditional methods. Continuous optimization efforts are made to refine the models and techniques, aiming to achieve the highest levels of comfort and energy efficiency.

By adopting this methodological approach, the research aims to provide a validated, holistic framework for building design that integrates comfort, efficiency, and sustainability. The ultimate goal is to contribute to the creation of more efficient, comfortable, and sustainable living and working environments through innovative building design practices.

1.6 Research Contribution

This research makes significant strides in building design, particularly enhancing comfort and energy efficiency through the application of machine learning techniques. A key contribution is the development and validation of a comprehensive comfort design framework. Supported by machine learning, this framework integrates various comfort aspects such as thermal, acoustic, visual comfort, and air quality into a unified approach. It aims to create buildings that are not only energy-efficient but also promote occupant well-being, marking a substantial leap towards holistic building design.

Addressing the need for a holistic approach in the industry, this research introduces a standardized normalization method for quantifying and assessing building comfort. By providing a unified total building comfort index, it enables designers and engineers to objectively measure and compare comfort levels across different buildings and designs. This standardization is a pivotal contribution, offering a quantifiable means to assess and enhance comfort in building environments.

The research also significantly advances the predictive capabilities in building design through the development of machine learning models for energy consumption and occupant comfort. These models represent a significant improvement in accuracy and efficiency over traditional methods, providing precise control over building performance and enabling the optimization of energy usage.

Integrating machine learning with Building Information Modelling (BIM) is another notable contribution, enhancing the data-driven design process. This integration allows for the effective leveraging of extensive BIM data, facilitating better design decisions and outcomes. Additionally, the research offers practical guidelines and best practices for incorporating machine learning into building design, providing valuable insights into the challenges and opportunities presented by these advanced techniques.

Furthermore, by focusing on energy efficiency and comfort optimization, the research contributes to more sustainable building practices. It supports the broader goal of reducing the environmental impact of buildings and enhancing the quality of life for occupants. This research also addresses a significant gap in the literature by focusing on comprehensive comfort in building design, providing valuable insights and empirical evidence on the effectiveness of machine learning in creating holistic and adaptive building environments.

Through these various contributions, the research aims to influence both academic and industry practices, pushing the boundaries of building design and operation. It sets a new standard for how buildings are designed, constructed, and operated, with a clear focus on energy efficiency, occupant comfort, and overall sustainability.

1.7 Structure of Dissertation

This dissertation is organized into six chapters, described in detail as follows:

Chapter 1: Introduction. This chapter presents the background, problem statement, research objectives, and hypotheses of the study. It sets the scope and motivation for the research, providing the reader with a fundamental understanding and necessity of the study.

Chapter 2: Literature Review. This chapter reviews existing literature in relevant fields, including building comfort, energy efficiency, applications of machine learning, and the theoretical and practical advancements related to these areas. It aims to reveal the research gaps and the niches that this study intends to fill.

Chapter 3: Methodology. It details the methods and techniques employed in the research, including data collection, development and training of machine learning models, standardization and quantification of comfort measures, and the overall logic and structure of the research design.

Chapter 4: Establishment of Comprehensive Comfort Framework. This chapter elaborates on the construction of a comprehensive comfort framework, including the theoretical basis of the framework, the steps in building it, and how different comfort indicators are integrated. The normalised building comfort function is built in this chapter.

Chapter 5: Application of Machine Learning in the Comfort Framework. It introduces how machine learning techniques are applied within the comprehensive comfort framework, including the selection, training, testing, and optimization of models. This chapter also demonstrates how machine learning improves the accuracy and efficiency of comfort predictions and discusses case studies of its application in different building designs from real buildings.

Chapter 6: Development of Customized Ontology for Comfort Assessment. This chapter introduces the ontology-based approach used to integrate and manage the various factors influencing building comfort. This chapter highlights the role of Semantic Web Rule Language (SWRL) and how it facilitates logical reasoning within the system.

Chapter 7: Conclusion and future works. This chapter summarizes the dissertation, revisiting the main findings, contributions, and implications for practice. It discusses the limitations of the research and potential directions for future research, providing insights and suggestions for subsequent studies.

Through this six-chapter structure, the dissertation aims to provide a systematic and comprehensive account of the research process, addressing the posed research questions while offering in-depth insights and solutions for sustainability and comfort in building design and operation.

2 Literature Review

2.1 Building Information Modelling (BIM)

BIM, an abbreviation for Building Information Modelling, is a method used to monitor the entire life cycle of a construction project (Eleftheriadis, Mumovic, and Greening, 2017). BIM captures multi-dimensional Computer-Aided Design (CAD) information (Eadie et al., 2013). This digital revolution has significantly contributed to the advancement of the Architecture, Construction, and Engineering (ACE) industry. Data can be easily accumulated and collected through integrated BIM models, which embed comprehensive building information. According to Krygiel, Nies, and McDowell (2008), the concept of BIM refers to an integrated database that stores all parametric and interconnected information of an entire building, as well as design documents. Any changes made to an object in the model are instantly reflected across the project in all views.

The terms "green building" and "sustainable design" have become common in recent years. However, many people confuse the difference between these two concepts. It is said that green buildings have a smaller environmental impact compared to traditional buildings, though the extent of this impact remains unclear (Krygiel, Nies, and McDowell, 2008). Sustainable design, on the other hand, is a more advanced concept in the industry, considering broader impacts across the entire life cycle of a building. The most precise definition of sustainable design was provided by the World Commission on Environment and Development in 1987 (WCED, 1987). Sustainable development refers to meeting the needs of the present without compromising the ability of future generations to meet their own needs. Sustainable design is crucial because it focuses on three key factors: people, planet, and prosperity (Krygiel, Nies, and McDowell, 2008). Firstly, many building materials can cause hazards to occupants or construction workers, such as formaldehyde is one the common chemical carcinogens realised by the building materials (Pacheco-Torgal, 2012). So, materials that pose long-term health risks to occupants should be eliminated. Secondly, the environmental impact of industrial processes has been a concern for decades. Lastly, sustainable design can reduce the costs associated with green buildings, resulting in a shorter return on investment period. There are five rating systems for green buildings: CASBEE, SBTool, BREEAM, Green Globes U.S., and LEED. BREEAM is the earliest of the five, having been introduced in 1990s, and it has since been widely used in the UK (Prior, 1993).

BIM serves as a communication platform among project stakeholders, facilitating improved efficiency (Krygiel, Nies, and McDowell, 2008). BIM encompasses a wide range of models, including the physical design model, solar analysis model, digital design model, energy model, daylight model, and construction documents model. BIM integrates these various models into a single, comprehensive model. The energy model within BIM not only displays energy load and demand but also illustrates the building's integrated systems.

2.1.1 Machine Learning in the BIM area

The problems addressed by BIM and machine learning share similarities, as BIM employs a top-down approach to modelling information. In contrast, machine learning (ML) adopts a data-driven, bottom-up approach to identify structure and semantics in data. Therefore, ML 'learns' from data, whereas BIM facilitates 'knowledge discovery' by providing information directly from the model. Although numerous studies have explored the application of machine learning in building design, there is limited research on leveraging BIM for knowledge discovery.

The evolution from CAD to BIM can be mapped onto the data-information-knowledge (DIK) hierarchy, where CAD models represent pure data, while Building Information Models (BIM), as implied by the term, are considered information (Figure 2.1). BIM is semantically rich; BIM stands in relation to CAD as LaTeX does to regular text (e.g., in PDF) or as the Semantic Web does to the traditional Web. This distinction will become clearer through some definitions.

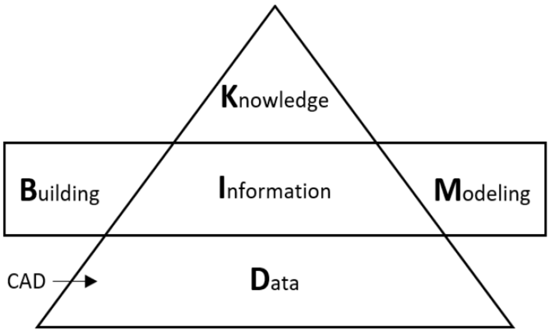


Figure 2. 1 CAD to BIM evolution projected on DIK pyramid（Foux, 2019）

Knowledge is not inherently embedded within BIM models; however, BIM provides a foundation for knowledge discovery. Building performance indicators, such as energy consumption, daylight utilisation, thermal and visual comfort, air quality, and environmental impacts, are examples of knowledge that can be derived from a design solution. Another set of examples pertains to design quality in terms of procurement, constructability, and operation, including factors such as cost to build, maintain, and operate, clashes between elements from different disciplines, and the availability of materials on the market.

As shown in Table 2.1 below, the applications of machine learning in BIM can be categorised into two main classes: (1) As-constructed BIM, which transforms data into information (BIM models), and (2) Implicit Information and Knowledge.

Table 2. 1 Applications and implications of machine learning are used in BIM

Classification	Example	Function
As-constructed BIM (transform data to information (BIM models))	1 Automated generation of as-built BIM models based on laser scanning data usually represented by point cloud (Tang et al., 2010).	Automated construction progress monitoring when combined with 4D BIM
	2 Utilizes data from daily construction photo collections (instead of laser scanning data) and structure from motion techniques to generate geometry point clouds (Golparvar-Fard et al. 2011).	Automatically detect physical progress in the presence of occlusions.
Implicit Information and Knowledge	1 Detect anomalies/outliers in a BIM model (Abouelaziz et al.2023).	Automatic conflict detection and quantity take-off, followed by material procurement in the construction phase.
	2 Classify system (Ryu, 2020).	Automatically assign missing attributes, use in archiving systems

2.1.2 Information Extraction from BIM

Information extraction from Building Information Modelling (BIM) involves the process of retrieving and organizing relevant data embedded within the BIM models. BIM models contain vast amounts of information related to various aspects of a building, such as geometry, spatial relationships, building components, materials, and performance metrics. Extracting this data efficiently is essential for various stakeholders, including architects, engineers, and project managers, to inform decision-making throughout the design, construction, and operational phases of a building's lifecycle.

The extraction process typically relies on specialized software tools capable of navigating through the BIM models to retrieve specific data points. These tools can extract information such as:

- **Geometrical data**: 3D shapes, dimensions, and spatial relationships between different building elements.
- **Material data**: Types of materials used in construction, their properties, and sources.
- **Performance data**: Energy consumption, lighting efficiency, thermal comfort, and air quality metrics.
- **Cost and scheduling data**: Information about the project timeline, estimated costs, and procurement schedules.

Moreover, information extraction from BIM facilitates advanced analysis, such as clash detection, cost estimation, sustainability assessment, and building performance optimization. By extracting and organizing this data effectively, BIM enhances collaboration between different stakeholders and improves the overall efficiency of the construction process.

The creation of IFC files begins with the modelling of a building project in specific BIM software by stakeholders involved in the project (Dhillon, 2014). Since IFC is not a native file format for any building modelling software, all building information models are first created in other software packages, such as Autodesk Revit, Bentley MicroStation, Graphisoft ArchiCAD, or Nemetschek VectorWorks. For the modelling process, objects are assembled from a library that typically includes objects provided by the software company, objects supplied by manufacturers of building components, and custom objects created by users for specific purposes (ibid). These objects consist of 3D geometry (boundary representation), property data, and metadata. This information is retained when the model is exported to the IFC format. However, other information, particularly 2D representations of objects, is not included in the IFC format and is therefore lost during the export from the native file format.

IFC (Industry Foundation Classes) is a standard for building information models, rather than for traditional 2D drawings (Borrmann, 2018). At a general level, it enables users to exchange information about building structures, elements, spaces, and other objects within a Building Information Model. IFC facilitates this information exchange through several key methods(ibid):

- Information Delivery Manual (IDM): A methodology designed to capture and specify processes and information flow throughout the lifecycle of a facility.
- Model View Definition (MVD): Defines a subset of the IFC schema required to satisfy one or more Exchange Requirements within the Architecture, Engineering, and Construction (AEC) industry. Examples include Coordination View, Structural Analysis View, and FM Handover View (COBie).

The content of an IFC file is determined by the specific MVD applied, which defines its scope(Borrmann, 2018). This content typically includes 3D geometry, properties, and attributes (such as parameters, relationships, and connectivity). The objects in an IFC file are aware of the systems they belong to and their connections to other objects. Ideally, an IFC file should include all necessary information to understand its content on both a qualitative level (e.g., author, location) and a quantitative level (e.g., measurements, quantities). Where certain information is not explicitly embedded in the file, algorithms can derive it through methods such as counting elements or measuring areas. The IFC schema includes hundreds of predefined properties, which are populated by 3D software during export from the native format(ibid).

2.2 Application of Machine Learning in the Context of Big Data

2.2.1 Introduction to Machine Learning

Data is omnipresent in all aspects of human life around the world. Machine learning has become a popular method for mining information and knowledge from data (Witten et al., 2016). The dictionary definition of "learning" is "to acquire knowledge by study, experience, or being taught." However, machine learning differs from human and animal learning, although it simulates the functioning of the human/animal brain (ibid). The concept of machine learning is inspired by how humans and animals learn from natural experiences.

Machine learning utilizes past data or example data through programming, without relying on a predetermined equation, to optimize a performance criterion. A formal definition of machine learning, provided by Mitchell (1997), states that the hypothesis P estimates a computer's performance in task T. A computer program is said to learn from experience E when its performance in task T improves using experience E. Alpaydin (2014) further defines machine learning as "programming computers to optimize a performance criterion using example data or past experience." Thus, the simplest and most direct meaning of machine learning is the use of a computer program to simulate a person's ability to learn implicit knowledge from real-world instances.

The application of machine learning to large-scale databases is referred to as data mining (Alpaydin, 2014). However, machine learning is not solely a database problem; it is also an integral part of artificial intelligence (AI). Machine learning is a branch of AI and one of its key implementation methods, used for practical inference and decision-making. A significant distinction between machine learning and traditional programming is its reliance on sample data, making it a data-driven approach.

2.2.2 Development History of Machine Learning

The development of artificial intelligence (AI) has progressed through three distinct stages: logical reasoning, knowledge engineering, and Machine learning (Hopgood, 2021). The first stage focuses on logical reasoning, such as proving mathematical theorems, which involves using symbolic logic to simulate human intelligence (Genesereth, 2012). The second stage is characterized by the advent of expert systems. These systems build expert knowledge bases for problems in various fields and use this knowledge to perform reasoning and decision-making (Hunt, 2012). For instance, if AI is applied to disease diagnosis, it requires constructing a database of doctors' diagnostic knowledge, which is then used to evaluate patients. The third and current stage of AI development is machine learning.

Although the term "machine learning" and some of its foundational methods can be traced back to 1958 or even earlier, the true emergence of machine learning (ML) as an independent discipline began in the 1980s, marked by the first academic conferences and journals dedicated to the field (Kodratoff, 2014). Since then, the development of machine learning has gone through three distinct stages:

- **First stage (1980s):** This period saw the formalization of machine learning as a discipline, but it remained relatively non-influential, with limited practical applications.
- **Second stage (1990s–2010s):** During this period, many foundational theories and algorithms were developed and began to be applied in practical contexts, significantly advancing the field.
- **Third stage (post-2012)**: The advent of deep learning, which marked the beginning of the third stage, led to significant breakthroughs in solving key artificial intelligence problems. The rapid development of deep learning has not only advanced the field of AI but also catalyzed rapid growth in various industries.

The timeline of major machine learning (ML) algorithms over the past decades is illustrated in Figure 2.2 . In 1958, logistic regression was proposed. Logistic regression fits a concise model to predict the probability that a binary response belongs to one category or the other. As a result, logistic regression is commonly used in binary classification problems where the data can be clearly separated by a single, linear boundary (Alpaydin, 2014). Subsequently, the k-Nearest Neighbour (kNN) algorithm was introduced, which classifies objects based on the class of the nearest neighbours. The underlying hypothesis of kNN is that objects near each other are likely to be similar.

In 1980, machine learning emerged as an independent discipline. Over the following decade, several important methods and theories were developed, including the Classification and Regression Tree (CART) in 1984, the backpropagation algorithm in 1986, and the convolutional neural network (CNN) in 1989. Decision trees for regression are similar to decision trees for classification but are adapted to predict continuous responses. These are particularly useful when predictors are categorical (discrete) or exhibit nonlinear behaviour.

Artificial neural networks (ANNs) are a simple simulation of the animal nervous system and belong to the field of bio-inspired methods. Neural networks cover a wide range of approaches within this field. According to Kohonen (1988), neural networks are composed of interconnected simple neurons/units, which can mimic the complex reactions of living organisms to stimuli in the real world. The most basic element in neural networks is the neuron model, also referred to as a simple unit. In biological neural networks, neurons are interconnected. When a neuron is activated by external stimuli, it releases chemicals that alter the electrical potential in neighbouring neurons. If the electrical potential of a neuron exceeds a certain threshold (also known as bias), it becomes activated and, in turn, releases chemicals to influence other neurons.

The backpropagation algorithm is a method for training neural networks, making multi-layer neural networks a practical and valuable machine learning methodology. Convolutional neural networks (CNNs) are a class of neural networks inspired by the principles of the animal visual nervous system, and they were originally modelled to abstract and understand images. CNNs were first applied to recognize handwritten digits by LeCun at Bell Laboratories.

From 1990 to 2012, machine learning matured significantly, with numerous theories and methodologies emerging, such as Support Vector Machine (SVM) (1995), AdaBoost (1997), Recurrent Neural Network (RNN) (1997), Long Short-Term Memory (LSTM) (1997), and Random Forest (2001). SVM is based on maximizing the classification margin, which is formulated as a convex optimization problem, providing excellent generalization capabilities. AdaBoost and Random Forest are both boosting algorithms that enhance performance by combining multiple weak learning models into a stronger model. RNNs are feedforward neural networks with internal states (memory), enabling them to process sequences of inputs. Despite the success of neural networks, they initially suffered from poor training performance and the problem of converging to local optima. In 2006, Hinton and his team developed a new, efficient method for training deep neural networks. This method was successfully applied in 2012 to AlexNet, a deep convolutional neural network, which achieved significant performance improvements. Since then, deep neural networks have been widely applied in reinforcement learning (RL), such as in AlphaGo, and in Generative Adversarial Networks (GANs), which are used for data generation.

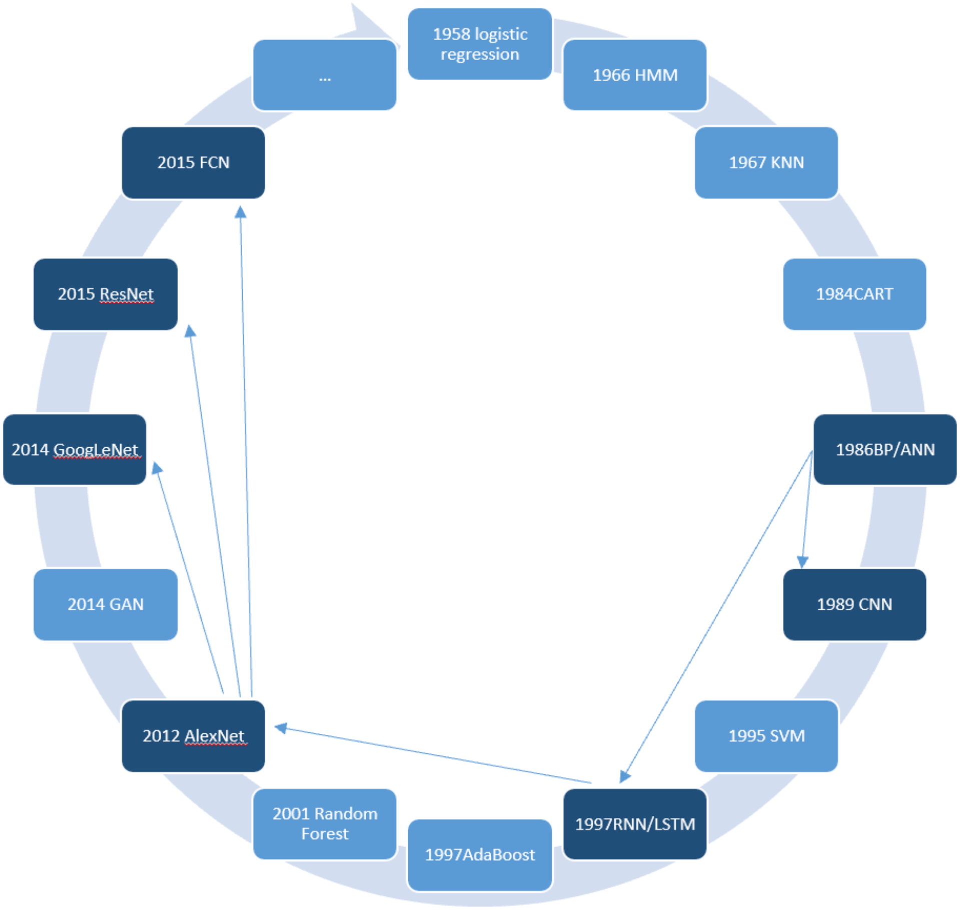


Figure 2. 2 The Timeline of Main ML Algorithms in Recent Decades

2.2.3 Machine Learning Algorithms

Machine learning encompasses a wide variety of algorithms, which can be broadly divided into two main categories based on label type: supervised learning and unsupervised learning. Supervised learning refers to algorithms that learn from labelled data, where the model is trained on input-output pairs to make predictions (Jain *et al*., 1999). In contrast, unsupervised learning deals with unlabelled data, where the model identifies patterns or structures within the data without explicit outputs. Figure 2.3 below illustrates the general categories of machine learning algorithms.

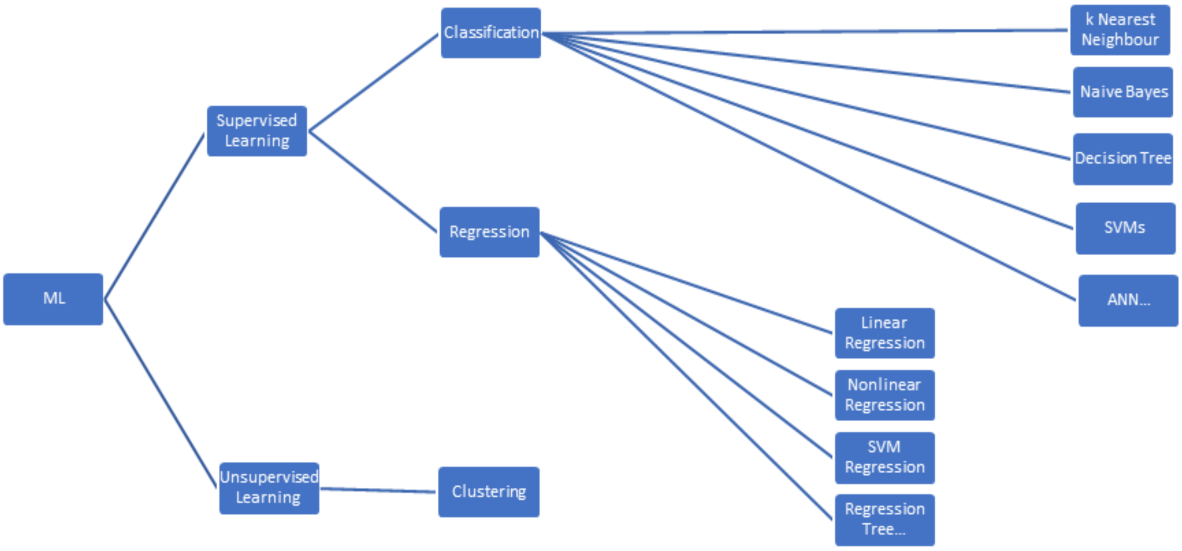


Figure 2. 3 General ML Algorithms

- Supervise Learning

Supervised learning refers to the scenario where instances are provided with known labels, corresponding to correct outputs. In contrast, when instances are unlabelled, the process is referred to as unsupervised learning (Jain et al., 1999). Supervised learning leverages known input and output data to generate a concise model that maps inputs to class labels or predicts continuous outputs (Kotsiantis et al., 2006). Both classification and regression techniques are commonly used in supervised learning to train predictive models. In other words, all supervised learning methods are either forms of classification or regression.

Classification techniques focus on discrete outcomes, such as determining whether an email is genuine or spam, or whether a tumour is cancerous or benign, which can be classified into different categories. Common applications include medical imaging, speech recognition, and credit scoring. Several algorithms rely on classification techniques, such as Support Vector Machines (SVM), Discriminant Analysis, Naive Bayes, and k-Nearest Neighbours (kNN). Selecting the appropriate algorithm is a critical step in developing a classifier. The evaluation of classifiers is primarily based on their prediction accuracy.

There are three main methods used to calculate the accuracy of a classifier:

- Train-Test Split: The training set is divided into two groups—two-thirds is used for training, while the remaining one-third is used to estimate the model's performance.

- Cross-Validation: The training set is split manually into equal-sized subsets, and the classifier is iteratively trained on one subset while using the others for validation. This technique provides a robust estimate of performance.
- Leave-One-Out Validation: This is a special case of cross-validation, where each instance in the dataset is used once as a test set, while the remaining data is used for training. Though computationally expensive, it is useful when the most accurate estimate of a classifier's error rate is required.

Here are Some Common classification algorithms listed in Table 2.2 below (Alpaydin, 2014).

Table 2. 2 Common Classification Algorithms

Naive Bayes	Naive Bayes assumes the presence of a particular feature in a class in unrelated to the presence of any other feature. It classifies new objects based on the highest probability of belonging to a particular class. It often suits in small dataset containing many parameters. When dealing with text, it is often to treat each unique word as a feature. Naive Bayes performed well in text classification because of its simplicity and independent features. However, Naive Bayes is not considering the order of the words.
Discriminant Analysis	Discriminant analysis classifies data by finding linear combinations of features. It assumes that different classes generate data based on Gaussian distributions. Training a discriminant analysis model involves finding the parameters for a Gaussian distribution for each class. The distribution parameters are used to calculate boundaries, which can be linear or quadratic functions. These boundaries are used to determine the class of new data.

Decision Tree	A decision tree predicts responses to data by following a series of decisions from the root node (the starting point) down to a leaf node (the final outcome). The tree is composed of branching conditions, where the value of a predictor is compared to a threshold or weight learned during the training process. The number of branches and the values of these weights are determined during training. To improve the model's generalization and prevent overfitting, additional techniques such as pruning may be applied to simplify the tree by removing unnecessary branches.
Support Vector Machines (SVMs)	In machine learning, Support Vector Machines (SVMs, also known as support vector networks) are used for classification and regression tasks within supervised learning. Given a training set of examples, each labelled as belonging to one of two categories, the SVM algorithm creates a model that assigns new examples to one of these categories, functioning as a non-probabilistic binary linear classifier. The SVM model can be represented by points in a multi-dimensional space, where a hyperplane separates the data points of different categories. New data points are mapped into the same space, and their classification is determined by which side of the hyperplane they fall on. When the data is linearly separable, the optimal hyperplane is the one that maximizes the margin between the two classes. For non-linearly separable data, a loss function is introduced to adjust the classification of points that fall on the wrong side of the hyperplane.
Artificial neural network (ANN)	Neural networks are inspired by the human brain, where neurons are interconnected. When a neuron is stimulated, synapses release a chemical substance called neurotransmitter, which transmits signals through the neural network to the brain. Similarly, an artificial neural network (ANN) consists of highly connected layers of artificial neurons that relate inputs to desired outputs. The network is trained by iteratively adjusting the strengths of the connections (weights) so that given inputs are correctly mapped to the appropriate outputs. ANNs are particularly effective in modelling highly nonlinear systems.

	The concept of ANNs was introduced over 60 years ago, but their practical application has been prominent only in the last 30 years. ANNs have been applied in various fields, such as mathematics, engineering, medicine, economics, meteorology, psychology, neurology, and more. Some of the most important contributions of ANNs include pattern recognition, sound and speech recognition, electromyogram analysis, medical diagnostics, military target identification, and detecting explosives in passenger luggage. Additionally, ANNs are widely used for weather forecasting and market trend prediction. One of the key advantages of ANNs is their ability to perform successfully in areas where other methods have failed, particularly in complex, nonlinear systems. Secondly, ANNs have the ability to analyse large and complex systems, making them particularly useful in environments where the relationships between variables are highly nonlinear and difficult to model using traditional methods. Lastly, they are highly fault-tolerant, robust, and resistant to noise, allowing them to maintain performance even when input data is incomplete or contains errors (Kalogirou, 2018).

Meanwhile, the common regression algorithms are listed in Table 2.3 below.

Table 2. 3 Common Regression Algorithms

Linear Regression	Linear regression is a statistical modelling technique used to describe a continuous response variable as a linear function of one or more predictor variables. Because linear regression models are simple to interpret and easy to train, they are often the first model to be fitted to a new dataset. It is also can be used as a baseline for evaluating others.
Nonlinear Regression	Nonlinear regression is a statistical modelling technique that helps describe nonlinear relationships in experimental data. Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation. “Nonlinear” refers to a fit function that is a nonlinear function of the parameters. For example, if the fitting parameters are b_0, b_1, and b_2: the

	equation $y = b_0+b_1x+b_2x_2$ is a linear function of the fitting parameters, whereas $y = (b_0xb_1)/(x+b_2)$ is a nonlinear function of the fitting parameters. It can be used in a situation when data has strong nonlinear trends and cannot be easily transformed into a linear space.
Gaussian Process Regression Model	How it Works Gaussian process regression (GPR) models are nonparametric models that are used for predicting the value of a continuous response variable. They are widely used in the field of spatial analysis for interpolation in the presence of uncertainty. GPR is also referred to as Kriging. It is suit for interpolating spatial data, such as hydrogeological data for the distribution of ground water. Also, it can be used as a surrogate model to facilitate optimization of complex designs such as automotive engines.
SVR	SVR algorithms work like SVM classification algorithms but are modified to be able to predict a continuous response. Instead of finding a hyperplane that separates data, SVM regression algorithms find a model that deviates from the measured data by a value no greater than a small amount, with parameter values that are as small as possible (to minimize sensitivity to error). It is suit for high-dimensional data (where there will be a large number of predictor variables)
Generalized Linear Model	A generalized linear model is a special case of nonlinear models that uses linear methods. It involves fitting a linear combination of the inputs to a nonlinear function (the link function) of the outputs. It is suit for that the response variables have nonnormal distributions, such as a response variable that is always expected to be positive
Regression Tree	How It Works Decision trees for regression are similar to decision trees for classification, but they are modified to be able to predict continuous responses. It is commonly used that predictors are categorical (discrete) or behave nonlinearly

- Unsupervised Learning

Unsupervised learning is useful when the goal is unclear or when the underlying patterns in the data are not known. It can also be employed to reduce the dimensionality of data, making it easier to interpret or analyse. Most unsupervised learning techniques are a form of cluster analysis, where data is grouped based on similarity or shared characteristics (Hastie, 2009). In cluster analysis, objects within the same cluster are similar, while objects in different clusters are distinct.

Clustering algorithms can be broadly categorized into two groups: hard clustering and soft clustering (ibid). In hard clustering, each data point belongs to exactly one cluster, whereas in soft clustering, each data point can belong to multiple clusters with varying degrees of membership. Table 2.4 below provides an introduction to some common unsupervised learning algorithms.

Table 2. 4 Unsupervised learning algorithms

Common Hard Clustering Algorithms	*k*-Means	*k*-Means algorithm is that data is partitioned into k number of mutually exclusive clusters. How well a point fits into a cluster relies on the distance from that point to the cluster's centre. It is commonly used when the number of clusters is already known. Also, it suits the large data sets clustering.
	k-Medoids	*k*-Medoids is similar to k-means, but with the requirement that the cluster centres coincide with points in the data.
	Hierarchical Clustering	When the number of clusters is unknown, and the selection process can be visualization by the Hierarchical Clustering algorithm. It produces nested sets of clusters by analysing similarities between pairs of points and grouping objects into a binary, hierarchical tree (Hastie, 2009).
	Self-Organizing Map	Self-Organizing Map is a neural-network based clustering that transforms a dataset into a topology-preserving 2D map. It not only can visualize high-dimensional data in 2D or 3D, but also can deduce the dimensionality of data by preserving its topology (shape).

Common Soft Clustering Algorithms	Fuzzy c-Means	Data partition-based clustering when data points may belong to more than one cluster. It is suit for pattern recognition when the number of clusters is known but overlap.
	Gaussian Mixture Model	Partition-based clustering is where data points come from different multivariate normal distributions with certain probabilities. It is best used when a data point might belong to more than one cluster, or the clusters have different sizes and correlation structures within them. The model of Gaussian distributions that gives probabilities of a point being in a cluster.

2.2.4 Deep Learning

Deep neural networks (deep learning) are an outstanding component of machine learning, excelling in the domains of pattern recognition. A typical neural network, which is analogous to the human brain, consists of multiple simple processing units called neurons, which are interconnected by weighted connections (Schmidhuber, 2015). When a neuron receives stimulation, synapses release a chemical substance known as a neurotransmitter. This signal is transmitted through the neural network, similar to how signals are transmitted in the human brain. Theoretically, the more parameters a model has, the more expressive it becomes, and the greater its capacity to solve complex learning tasks. However, in practice, more complex models often suffer from lower training efficiency and are prone to overfitting (Srivastava, 2014). For this reason, such training methods were not widely adopted in previous decades. However, with the development of cloud computing and big data, advancements in computational power have significantly improved training efficiency. Additionally, the availability of larger datasets helps reduce the likelihood of overfitting. As a result, deep learning has become highly popular today.

There are three popular neural network architectures widely used in the field of deep learning.

- Convolutional Neural Networks (CNNs) are highly effective in image recognition, video analysis, natural language processing, and chemical recognition (LeCun, Bengio, and Hinton, 2015). CNNs offer several advantages, such as their ability to process pixels directly to recognize images in the domain of photo recognition. However, CNNs have certain limitations, such as the loss of information during the training process due to the use of strides. To address this, pooling layers are introduced to compress the information and mitigate the issue of information loss.

- Recurrent Neural Networks (RNNs) are commonly used in tasks like image description, script writing, and composing music (ibid). While RNNs are powerful, they also suffer from issues such as the vanishing gradient or exploding gradient problems during training. To address these challenges, the Long Short-Term Memory (LSTM) architecture is used, which helps retain information over long sequences and stabilizes the training process.
- Autoencoders are another type of neural network, which belong to the domain of unsupervised learning (ibid). In an autoencoder, image data is compressed by the encoder, and the compressed representation is compared to the original data to detect errors. The model then uses a Backpropagation Neural Network to improve its accuracy by reducing the error.

A case study in the construction industry highlights the use of optimization algorithms in energy management. Optimization algorithms are generally categorized into two main types: conventional gradient-based methods and gradient-free methods (Bejay, 2016). Artificial intelligence (AI) techniques have been increasingly applied in optimization algorithms for energy management (ibid). In the building domain, the focus is on the demand side, while the district domain represents the supply side. Both demand and supply sides must be considered when optimizing energy usage. Building energy optimization is complex and influenced by multiple factors. The goals of multi-objective optimization often include minimizing energy consumption and CO2 emissions while maximizing comfort levels. A significant amount of prior work has been conducted in this area. For example, Hamdy et al. (2011) used MATLAB in conjunction with a multi-objective genetic algorithm (GA) involving 24 design variables to identify energy-saving strategies. Similarly, Wang et al. (2005) employed a GA in an object-oriented framework to support simulation-based green building design. When comparing Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs), SVMs have the advantage of requiring less training data than ANNs. However, ANNs offer the benefit of faster training speeds compared to SVMs (Bejay, 2016).

There are several methods to evaluate the performance of a neural network. One common approach is to randomly divide the dataset into two parts: a larger portion for training the model and a smaller portion for testing it (LeCun, Bengio, and Hinton, 2015). The training data is used to develop the neural network, while the test data is used to evaluate its accuracy and performance.

2.2.5 Software/Platform/System of Machine Learning

Here are some related Software/Platform/Systems of machine learning listed below.

- Weka

Weka is a workbench for machine learning that is intended to aid in the application of machine learning techniques to a variety of real-world problems (Holmes, 1994). It is different from machine learning projects in that the emphasis is on providing a working environment for the domain specialist rather than the machine learning expert. Weka is an easy-to-use toolkit of machine learning without coding, which is friendly to any field other than computers.

- KNIME

KNIME is short for the Konstanz Information Miner, which is a free and open-source data analytics and integration platform (Berthold et al., 2009). The Development of KNIME was started in January 2004 by a team of software engineers at the University of Konstanz as a proprietary product. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface and use of Java DataBase Connectivity allows the assembly of nodes blending different data sources, including pre-processing for modelling, data analysis and visualization without, or with only minimal, programming. To some extent as an advanced analytics tool KNIME can be considered a Statistical Analysis System alternative.

- Anaconda

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system conda. The Anaconda distribution is used by over 13 million users and includes more than 1400 popular data science packages suitable for Windows, Linux, and MacOS. The Detailed guidelines of anaconda can be found on web: https://www.anaconda.com/anaconda-community/.

- Machine Learning Libraries in Python

Here is the form which illustrates several basic machine-learning libraries:

Table 2. 5 Popular machine-learning libraries

Numpy	NumPy is the fundamental package for scientific computing with Python. It is mostly used for solving matrix problems.
Pandas	Pandas is the most popular machine learning library written in Python, for data manipulation and analysis
Matplotlib	Matplotlib, a great library for Data Visualization
SciKit-Learn	A library that provides a range of Supervised and Unsupervised Learning Algorithms. This library mainly focused on model building.

2.3 Machine learning Used in Building Energy Analysis

2.3.1 Building Energy Consumption

Buildings account for 40% of total energy consumption and 36% of CO2 emissions in European countries ('Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings', 2010). To achieve the goals of energy conservation and environmental impact reduction, it is essential to accurately predict building energy consumption to optimize building performance. Building energy analysis is complex, as both energy usage patterns and building types vary significantly. The primary forms of building energy usage include heating and cooling loads, hot water production, and electricity consumption, among others (Zhao and Magoulès, 2012). Common building types include offices, residential buildings, and industrial facilities. Building energy performance is influenced by various factors, such as weather conditions, building structure, materials used, occupancy behaviour, and heating, ventilation, and air-conditioning (HVAC) systems. Several methods have been employed to predict building energy consumption, which can be broadly classified into three categories: engineering methods, statistical methods, and artificial intelligence (AI) techniques. Additionally, various methods have been used for detecting and diagnosing issues related to building energy design.

In 2003, Krarti conducted a review of artificial intelligence (AI) methods applied in building energy systems (BES) (Krarti, 2003). Similarly, in 2010, Dounis conducted a similar review on AI methods used in the BES field (Dounis, 2010). An overview of building energy prediction methods was later written by Zhao et al. in 2012 (Zhao and Magoulès, 2012). The methods are primarily categorized into four groups as follows:

- **The engineering methods**

Engineering methods refer to the calculation of building energy consumption based on physical principles, which have been developed over the past fifty years. These methods can be divided into two categories: the detailed comprehensive method and the simplified method. The detailed comprehensive approach involves highly precise calculations for all aspects of a building, including both the internal and external environments. The ISO has established a standard for calculating energy consumption related to space cooling and heating in buildings (ISO 13790:2008). There are numerous software tools designed to estimate building energy efficiency and sustainability, such as DOE-2, EnergyPlus, BLAST, and ESP-r (Mohd-Nor and Grant, 2014). These tools are maintained by the U.S. Department of Energy (DOE). Despite the efficiency of these software tools, there are still barriers in practical applications. The input data must be treated as precise building and environmental parameters. Without accurate input data, the simulation results will lack accuracy. Additionally, these tools are often complex and require expert operation, which can reduce efficiency.

In 2009, a review by Al-Homoud highlighted two simplified methods: the degree-day method and the bin method (Al-Homoud, 2009). The degree-day method is limited in that it can only estimate the energy consumption of small buildings. In contrast, the bin method is applied to larger buildings, particularly those where energy generation is internal and independent of outdoor conditions, or where the relationship between indoor and outdoor loads and temperature is non-linear. Key factors such as weather conditions and building characteristics play a crucial role in determining building energy consumption. Weather parameters like temperature, humidity, solar radiation, and wind speed are variable factors that influence the energy usage of buildings. White and Reichmuth made predictions of building energy consumption using the mean of monthly temperatures, which demonstrated higher accuracy compared to standard methods that use heating and cooling degree days or temperature bins (Lin, Jan, and Liao, 2017).

- **Statistical methods regression**

Statistical methods regression models simply show the relationship between the energy consumption and energy index with influencing variables. Most Statistical methods regression ware been used in flowing situation, such as just using weather condition to make predictions of building energy, energy index perdition and key parameter of energy consumption. In the simple model, the statistical regression can be used to calculate the energy signature which is correlated energy use and weather parameters (Pfafferott, Herkel and Wapler, 2005). Regression supported Conditional Demand Analysis (CDA) was developed by Aydinalp-Koksal and Ugursal show same accuracy as engineering methods and neural networks, but it was easier to use (Aydinalp-Koksal and Ugursal, 2008). However, there are some limitations in CDA method which was lack of detail and flexibility.

- **Artificial intelligence (AI) methods**

Artificial Neural Networks (ANNs) are among the most popular AI methods used in building energy management (Bilal et al., 2016). ANNs are particularly effective at solving nonlinear and complex problems. They have been applied in numerous energy-related applications within buildings, including solar water heating systems, solar radiation estimation, electricity usage monitoring, indoor airflow distribution due to wind speed, energy consumption prediction, indoor air temperature control, and HVAC system analysis (ibid). Input data is one of the most critical factors in the ANN approach. Data can be obtained from four main sources: onsite measurement, surveys, billing records, and simulations (Zhao, 2014). Due to the noise inherent in raw data, various data preprocessing techniques have been developed by researchers to enhance the quality and reliability of inputs. Numerous studies have compared the performance of ANNs with other methods in the field.

Support Vector Machines (SVMs) have also been applied to building energy information over the past decade, particularly in handling nonlinear problems. SVMs perform well with small quantities of training data. Previous studies have shown that SVMs are efficient for predicting building energy consumption on both hourly and monthly scales. However, there are limitations; for instance, most studies have applied this method to a small number of buildings to forecast future energy use. In 2010, Zhao et al. used SVMs to predict the heating load of multiple buildings, but the training process was extremely slow due to the large dataset size (Zhao and Magoulès, 2010). Subsequently, parallel SVMs were developed and used to improve training efficiency (Zhao and Magoulès, 2011). Additionally, Li et al. (2010) predicted electricity consumption in buildings using both general neural networks and SVMs.

- **Grey model**

There is also another model to do building energy analysis named the grey model, in which the information in the systems is partly known. Hence these methods can only be used in incomplete and uncertain data. Seldom works have been done before, such as Wang et al.(1999) use the grey model to predict the heat moisture system and it shows high accuracy

2.3.2 Application of ANNs in Building Energy

Kalogirou and Bojic (2000) showed the capabilities of ANNs used in building energy prediction. There are many difficulties in using artificial neural networks to analyse the building energy because of the nonlinear multivariate involved which needs to consider the inter-relationships between each other with noise. Additionally, the building energy system depends on the local weather conditions, for instance, solar radiation, wind speed, direction, strength and duration and so on. They have also been used in weather and market trends forecasting, in the prediction of mineral exploration sites, in electrical and thermal load prediction, in adaptive and robotic control and many others.

There are many applications of ANNs in energy application in buildings, which include the following: predicting solar radiation and wind, solar energy systems that can be applied in buildings, energy consumption prediction, energy conservation, HVAC system modelling and naturally ventilated buildings (Kalogirou, 1999). The details will be illustrated below.

- Solar Water Heating Systems

The first application of Artificial Neural Networks (ANNs) in this field was the prediction of a thermosiphon solar domestic water heating system (SDWH) (Kalogirou, 2018). A multi-layer feedforward neural network was used to train data collected from four types of systems under different weather conditions. The outputs of this model were the useful system energy and the change in stored water temperature. The maximum deviations for these two parameters were 1 MJ and +2.2°C, respectively. These results are highly consistent, demonstrating that ANNs can successfully predict thermosiphon performance under various conditions.

Another application involves the long-term performance forecasting of solar domestic water heating systems (Kalogirou and Panteliou, 2000). Thirty tests were conducted on thermosiphon SDWH systems at three different locations in Greece, following the ISO 9459-2 standard (ibid). The input data were divided into a training set (27 systems) and a testing set (3 systems). Monthly data were fed into two multi-layer feedforward neural networks to estimate the solar energy output of the systems for a draw-off quantity equal to the storage tank capacity and the average monthly quantity of hot water. The input data used were similar to those in the program provided by the standard, including system size, performance, and various climate data. Additionally, the second network incorporated the demand temperature as an input.

- Prediction of Solar Radiation

Solar radiation, including its intensity and availability, is influenced by many parameters due to its inherent nature. Therefore, multivariate prediction techniques are more suitable for predicting solar radiation.

In 1998, Al-Alawi and Al-Hinai used ANNs to predict solar radiation (Al-Alawi and Al-Hinai, 1998). The input data included location, month, mean pressure, mean temperature, mean vapour pressure, mean relative humidity, mean wind speed, and mean duration of sunshine. The model achieved an accuracy of 93% with a mean absolute percentage error (MAPE) of 7.3%. The result demonstrates the viability of this approach for spatial modelling of solar radiation.

In 1999, Kemmoku et al. employed a multistage ANN to forecast the next day's daily insolation (Kemmoku et al., 1999). The input data for this model consisted of the average atmospheric pressure, predicted by another ANN, along with various weather data from the previous day. The model achieved a prediction accuracy of 20%.

In 2002, Kalogirou et al. used ANNs to predict maximum solar radiation (Kalogirou, Michaelides, and Tymvios, 2002). The input data included factors that significantly influence the intensity and availability of solar radiation, such as month, day of the month, Julian day, season, mean ambient temperature, and mean relative humidity (RH). The model was trained using a standard backpropagation multilayer neural network, which is well-suited for time series prediction. Hourly records from an entire year were used as input to calculate the maximum radiation values and the mean daily values of temperature and RH. Data from 11 months were used for training, and the remaining month was used for testing. The experiment demonstrated a high level of accuracy.

In 2003, Reddy and Ranjan used an ANN to estimate the monthly mean daily and hourly values of global solar radiation (Reddy and Ranjan, 2003). They trained a multilayer feedforward perceptron to output the relative importance of a wide range of atmospheric and radiometric variables. The experiment indicated that this novel methodology can be effectively used under unfavorable conditions, such as when limited data is available, yielding successful results.

- Naturally Ventilated Buildings

Kalogirou et al. (1999) used Artificial Neural Networks (ANNs) to predict airflow in a single-sided, naturally ventilated test room. The test room in this experiment was equipped with adjustable louvers for ventilation. The researchers monitored local temperature, relative humidity, and wind velocity and direction, both indoors and outdoors. The dataset consisted of 32 trials, which were divided into two groups: 28 trials for training the model and 4 trials for testing. A multilayer feedforward neural network with three hidden layers was used. The results were satisfactory for predicting indoor temperature and combined airflow velocity when the model was tested with unseen data.

- Energy Consumption and Conservation

The process of heating load estimation in a previous study involved training ANNs to forecast building heating loads using minimal input data (Kalogirou, Neocleous, and Schizas, 1997). In this study, 250 known cases of heating load were used for training. The room types varied from small toilets to large classroom halls, with sizes ranging from 1 m² to 100 m². The room temperatures were controlled within a suitable range. Additionally, the characteristics of each room were incorporated. The input data included the areas of windows, walls, partitions, and floors, the types of windows and walls, whether the space had a roof or ceiling, and the designed room temperature. The output was the estimated heating load. The results of this study indicated that the accuracy of the prediction could be improved by grouping the input data into two categories: one for floor areas smaller than 7 m² and another for those larger than 7 m². The statistical value (R^2) was 0.9880 for the first category and 0.9990 for the second category. This study demonstrates that the proposed method can successfully predict building heating loads. Compared to conventional algorithmic methods, this approach offers three advantages: faster calculation speed, simplicity, and the neural network's ability to learn from examples, which allows for gradual improvement in performance.

Michalakou et al. (2002) used feedforward backpropagation neural networks to estimate residential building energy consumption in Athens. The input data included various climate parameters. The results were validated using extensive sets of non-training measurements, and they corresponded well with the actual values. In 2004, Aydinalp et al. used ANNs to predict residential end-use energy consumption at the national and regional levels. ANNs performed well in modelling the residential sector, including applications like lighting and space-cooling energy consumption. Two ANN energy consumption models were developed as a continuation of previous work: one for estimating space heating and another for domestic hot water heating.

- Heating, ventilation and air conditioning system

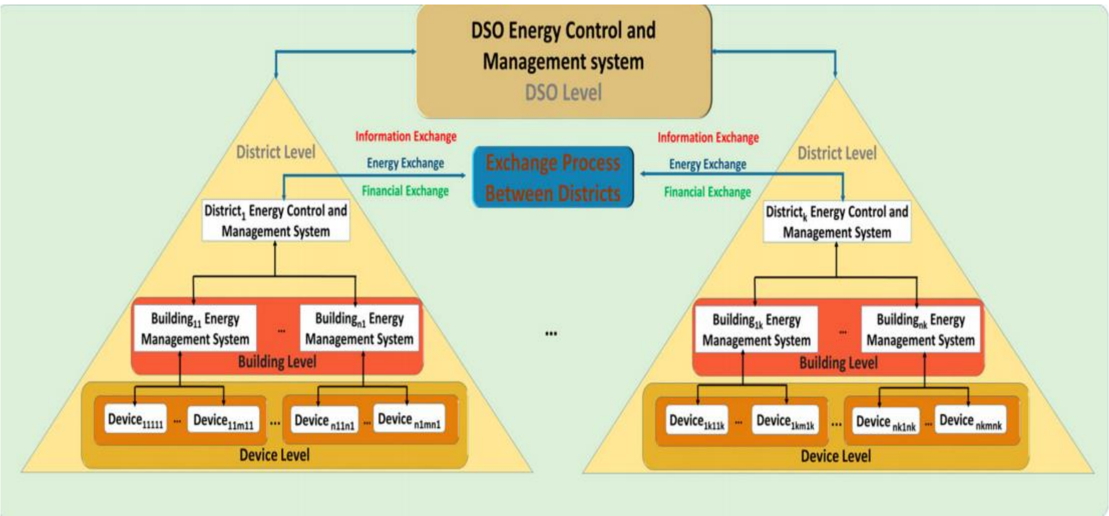
The model of discharge air temperature system is the first application in this field. In 2004, Zaheer-Uddin and Tudoroiu (2004) developed a nonlinear neuro-model of a discharge air temperature (DAT) system. The input data collected from the Heating, ventilation and air conditioning (HVAC) system test facility. The results indicate that 3 layers ANNs framework is performed good in predictions. Yang, Yeo and Kim (2003) used BPNN for optimization of building heating system through determining the star time. The input data are various building conditions which gather from the program simulation. Zmeureanu (2002) used the General Regression Neural Networks (GRNN) to propose a new method for estimation of the Coefficient of performance (COP) of existing rooftop units. This methods can lower the required monitoring devices, such as sensors, to decline the installation cost and recalibration or replacement costs during the operation.

2.3.3 ANN for Building Energy Analysis

Machine learning methods are increasingly popular in the load prediction domain due to their cost-effectiveness and ability to provide energy forecasts with finer temporal resolution. Kandananond (2011) applied three different methods to predict electricity demand in Thailand: the Autoregressive Integrated Moving Average (ARIMA) method, Artificial Neural Networks (ANN), and Multiple Linear Regression (MLR) using annual electricity consumption data.

Hernández et al. (2014) used solar radiation data as input for the ANN training engine to forecast energy generation on an hourly basis. The results indicated that the complexity of predicting electricity load for the next hour is related to disaggregated load prediction. In the same year, Idowu et al. (2014) used ANNs to forecast electricity demand at the district level, which can vary significantly due to household behaviour and financial circumstances. Peak demand is highly variable and complex, making it suitable for ANN applications. ANNs can also be utilized for peak demand determination. Given their high robustness and comparatively high accuracy, ANN-based models show potential for forecasting future building electricity demand at the district level.

A generalized smart grid energy management and control hierarchy, as illustrated in Figure 2.4 by Yuce, Mourshed, and Rezgui (2017), includes three hierarchical stages and one negotiation and exchange stage. The first stage is the device level, involving the activation and control of each device in the building. The second level is the building level, which focuses on the building's overall energy consumption. The third level is the district-level energy management, addressing the energy demand of a specific district, sometimes referred to as the aggregator energy management system. This system organizes the negotiation and exchange of information and finances between districts and connected buildings. The final stage pertains to the Distribution System Operator's (DSO) energy management system, which manages energy distribution among districts/aggregators. To optimize the entire process, predicting building-level energy demand becomes critical throughout the value chain. Consequently, an ANN-based forecasting system is proposed to predict the electricity demand of individual households within a selected district.



DSO Energy Control and Management system
DSO Level
District Level
Information Exchange
Energy Exchange
Financial Exchange
Exchange Process Between Districts
Information Exchange
Energy Exchange
Financial Exchange
District Level
District$_1$ Energy Control and Management System
District$_k$ Energy Control and Management System
Building$_{11}$ Energy Management System
...
Building$_{n1}$ Energy Management System
Building Level
...
Building$_{1k}$ Energy Management System
...
Building$_{nk}$ Energy Management System
Building Level
Device$_{11111}$ ... Device$_{11m11}$ ... Device $_{n11n1}$ ... Device $_{n1mn1}$
Device Level
Device$_{1k11k}$ ... Device$_{1km1k}$ ... Device $_{nk1nk}$ ... Device $_{nkmnk}$
Device Level

Figure 2. 4 Topology of district management in the smart grid, (Yuce, Mourshed and Rezgui, 2017)

A study on predicting the energy demand of commercial buildings during triad peaks in Great Britain was conducted at Cardiff University (Marmaras et al., 2017). The prediction process consisted of three stages, utilizing data from three different sources to enhance the accuracy of the proposed methods. The first stage involved a triad probability assessment model using 25 years of triad data gathered from the National Grid. The model aimed to determine the most likely dates and times of triad peaks for the forthcoming year. In the second stage, weather data collected from the Met Office was used in a pre-forecasting analysis model, which was an ANN built using the WEKA toolkit. The model performed best when the number of hidden layers matched the number of additional attributes, and when the number of neurons in each layer matched the total number of neurons. The third stage was the electricity demand forecast model. Five methods were compared: linear regression, instance-based learning, support vector regression (SVR), multi-layer perceptron (MLP) ANN, and decision tree. The ANN method was chosen for its superior performance compared to the other methods in the dataset. The input data included the most probable triad half-hours and the optimal forecast scenario, while the output was the predicted building power demand during a triad half-hour period.

2.4 Understanding Building Comfort

Thermal comfort refers to the psychological state of expressing satisfaction with the thermal environment, typically evaluated through subjective assessments. It is a feeling that can be challenging to define precisely. When converting this knowledge into machine-readable data, concerns about accuracy arise due to individual variability in thermal perception. While most people find room temperature, typically between 20-22°C, comfortable, this range can vary significantly between individuals and is influenced by factors such as activity level, clothing, and humidity. Personal factors greatly influence thermal comfort, but only environmental factors can be considered as general parameters for building design. However, certain personal factors may still be relevant for specific building uses.

The Predicted Mean Vote (PMV) model is one of the widely accepted thermal comfort models. It is based on the thermal equilibrium principle and experimental data collected under steady-state conditions in a controlled climate chamber. Another approach, the Adaptive Model, was developed based on hundreds of field studies, emphasizing the dynamic interaction between occupants and their environment. Occupants regulate their thermal environment using clothing adjustments, operable windows, fans, personal heaters, and sunshades. These two models will be explained in detail later。 Before that, the definitions of some related terms are provided below:

- Adaptive model

Adaptive modelling refers to the method of correlating interior design temperatures or acceptable temperature ranges with outdoor weather or climate parameters.

- Thermal comfort

Thermal comfort refers to the psychological state of expressing satisfaction with the thermal environment and is evaluated through subjective evaluation.

- Space of natural conditions under the control of the occupants

An occupant-controlled naturally regulated space is an opening in which the thermal conditions of the space are regulated primarily by the occupant.

- Projected mean value (PMV)

The predicted mean is the mean value of the number of heat sensation votes (self-reported sensations) predicted for a large group of people using a sensation scale corresponding to the categories "cold", "cool", "cool", "slightly cool", "neutral", "slightly warm", "warm", and "hot" in the range -3 to +3.

- Comfort zone

A comfort zone is a range of air temperature, mean radiant temperature (TR), and humidity that is predicted to provide an acceptable thermal environment under specific conditions of wind speed, metabolic rate, and clothing insulation (Icl). The insulating property of clothing (Icl) represents the resistance to heat transfer offered by a combination of garments, measured in clo units. One clo equals 0.155 m²·°C/W (0.88 ft²·h·°F/Btu), which quantifies the thermal insulation provided by clothing.

- Metabolic rate (met)

The metabolic rate is the rate at which an individual's metabolism converts chemical energy into heat and mechanical work per unit of skin surface area, expressed in "met." One met is equivalent to 58.2 W/m² (18.4 Btu/h-ft²) and represents the energy produced per unit of skin surface area while at rest.

- Excess hours

Excess hours refer to the total number of hours during a specified period when the environmental conditions in an occupied space exceed the limits of the comfort zone.

2.4.1 Thermal Comfort: Concepts and Measurement

This section delves into the factors influencing the thermal environment of a building, bifurcating them into external and internal elements. Externally, the building is affected by six primary factors: solar radiation, wind, outdoor air temperature, humidity, precipitation, and ground temperature. Solar radiation, quantified in W/m2, varies significantly with latitude, necessitating monthly adjustments for accuracy. Wind, driven by atmospheric pressure differences, includes both global circulation and localized breezes, characterized by its direction and speed. Outdoor air temperature, typically measured at 1.5m above ground in the shade, is influenced by solar radiation, ground cover, and atmospheric convection, exhibiting daily and seasonal variations. Humidity, both absolute and relative, significantly fluctuates throughout the day and is influenced by geographical and climatic factors. Precipitation, described in millimetres, is determined by temperature, atmospheric circulation, topography, and other factors. Lastly, ground temperature is influenced primarily by solar and long-wave radiation, with deeper strata temperatures remaining relatively constant.

Indoor microclimate is shaped by the building's enclosures (e.g., roof, walls, windows) and heating or air conditioning systems. Four key elements define this microclimate: temperature, humidity, air velocity, and thermal radiation. Temperature, particularly radiant heat, plays a crucial role in thermal comfort, emanating from various sources and significantly impacting human physiological activity and health. Humidity at moderate temperatures has negligible effects; however, extreme humidity levels can lead to discomfort and health risks. Air velocity affects thermal comfort differently across seasons and is vital for ensuring adequate ventilation. Thermal radiation involves heat exchange between the human body and its environment and is a critical factor in the sensation of warmth or coldness.

Furthermore, personal factors like clothing insulation and metabolic heat also play vital roles in determining thermal comfort. Clothing insulation is crucial in both preventing heat stress and providing adequate protection against cold, while metabolic heat reflects the physical activity level, influencing the body's heat production and dissipation.

By comprehensively analysing these diverse factors, this study aims to construct a nuanced understanding of thermal comfort, moving beyond conventional metrics to incorporate a wide array of environmental and personal influences. This approach seeks to enhance the precision and applicability of comfort assessment, contributing to the development of more habitable and energy-efficient buildings. These can be elements of purpose-specific construction.

- Clothing Insulation

Thermal comfort is significantly influenced by the insulation provided by clothing. Even when the environment is not particularly warm or hot, wearing excessive clothing or personal protective equipment (PPE) can lead to heat stress. Conversely, if the clothing does not offer sufficient insulation, the wearer may be vulnerable to cold-related injuries, such as frostbite or hypothermia. Clothing acts both as a potential source of thermal discomfort and as a means to regulate it, as individuals adapt to the climate in their work environment. For example, adding a layer of clothing when feeling cold or removing one when feeling warm is a common adaptation. However, many companies require employees to wear specific uniforms or PPE, which may limit their ability to adjust clothing for thermal comfort. It is crucial to evaluate the impact of clothing on thermal comfort. Regular assessments of the protection levels offered by existing PPE and evaluations of new PPE options can help improve thermal comfort in the workplace.

- Metabolic Heat (Metabolic Rate)

The more physical labour one performs, the more metabolic heat is produced. To prevent overheating, the body needs to dissipate this excess heat. The impact of metabolic rate on thermal comfort is crucial. When evaluating thermal comfort, it is important to consider individual physical characteristics, as factors such as body size, weight, age, fitness level, and gender can influence thermal sensation, even when environmental factors like air temperature, humidity, and wind speed remain constant.

2.4.2 Evaluation Methods of Thermal Comfort in a Building

These methods describe how thermal comfort is assessed in accordance with this standard and can be used as resources to calculate the comfort zone. The Predicted Mean Vote (PMV) model and the Standard Effective Temperature (SET) model are crucial in this evaluation.

The graphical comfort zone method utilizes a psychrometric chart to represent operational temperature and humidity levels for winter (1.0 clo) and summer (0.5 clo) to achieve thermal comfort. This method is based on the PMV model. Its applicability is limited when the occupant’s metabolic rate is between

1.0 and 1.3, and the humidity ratio is less than 0.012 kg H2O/kg dry air (0.012 lb H2O/lb dry air). If these criteria are met and the environmental conditions in the building are within the prescribed limits, then the comfort requirements are considered satisfied.

For humidity ratios exceeding 0.012 kg H2O/kg dry air (0.012 lb H2O/lb dry air) or metabolic rates up to 2.0 met, an analytical model must be used to determine thermal comfort. Also based on the PMV model, this method uses tools such as the ASHRAE Thermal Comfort Tool or the CBE Thermal Comfort Tool to assess comfort. The user inputs the operating temperature (or air temperature and mean radiant temperature), air velocity, humidity, metabolic rate, and clothing insulation value. The tool then evaluates the predicted thermal sensation on a scale from -3 (cold) to +3 (hot). If the conditions provide thermal neutrality, ranging between -0.5 and +0.5 on the PMV scale, the comfort requirements are considered met.

High Altitude Wind Speed This section provides for an increase in the upper air temperature limit when the high-altitude wind speed exceeds 0.20 m/s (39 ft/min). The method is based on the SET (Standard Effective Temperature) model, which provides a method of assigning effective temperatures (in the case of standard metabolic rates and clothing insulation values) to compare thermal sensations under a range of thermal conditions. The upper limit of air velocity is determined by whether the occupant has partial control. Localized thermal discomfort The radiant temperature asymmetry between the ceiling and the floor and between the air and the wall must be limited to reduce discomfort. In order to reduce the risk of ventilation at temperatures below 22.5 °C (72.5 °F), the air velocity caused by the HVAC system must be 0.15 m/s (30 ft/min) or less. The vertical air temperature difference between the ankles and the head shall not exceed 3°C (5.4°F) for sitting occupants and 4°C (7.2°F) for standing occupants. If the occupant's feet are in contact with the ground, the temperature must be 19-29 °C (66-84 °F). Temperature changes over time the conditions in this section must be met when the occupant is unable to control periodic changes or drifts in indoor environmental conditions. The operating temperature should not fluctuate more than 1.1°C (2.0°F) in 15 minutes and 2.2°C (4.0°F) in 1 hour. Acceptable thermal conditions under occupant-controlled natural conditions the method, also known as the adaptive comfort model, is applied to buildings without mechanical cooling (and without operating heating systems) where the occupants have a satisfaction rate of 1.0-1.3 m and a clothing level of 0.5-1.0 clo. For this model, the standard provides a diagram of acceptable indoor temperature limits at the then average outdoor temperature (the average of the average daily outdoor temperature for the first 7-30 days). An accompanying table lists the regulations for higher operating temperatures when the airspeed exceeds 0.3 m/s (59 ft/min) and up to 1.2 m/s (240 ft/min). The graph applies to a general average temperature between 10-33.5°C (50.0-92.3°F). It provides an acceptable range of 80% and 90%, showing the percentage of comfort at the specified average indoor and outdoor temperatures.

Here is mainly evaluation index of thermal comfort in Figure 2.5:

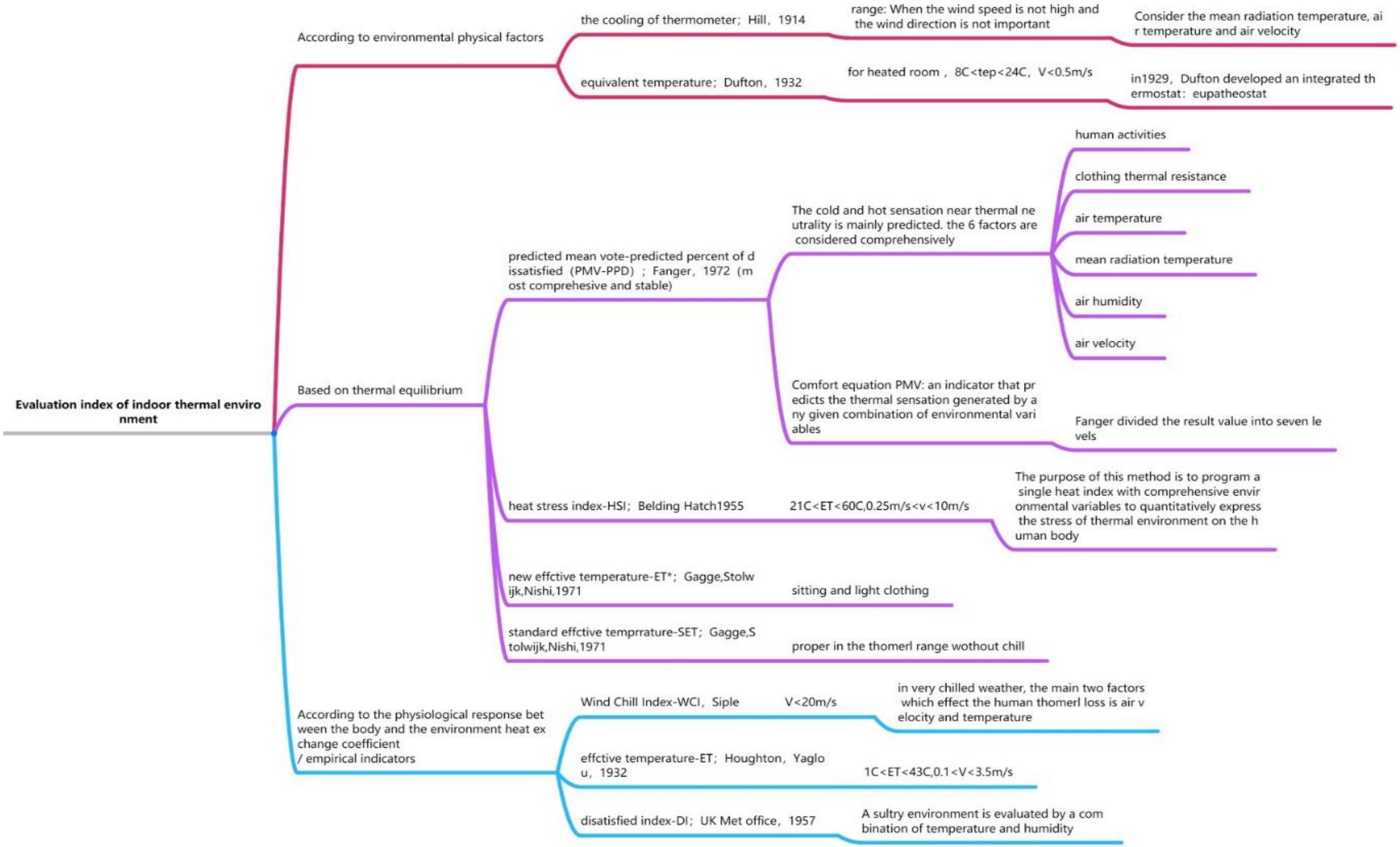

Evaluation index of indoor thermal enviro nment
According to environmental physical factors
the cooling of thermometer；Hill，1914
range: When the wind speed is not high and the wind direction is not important
Consider the mean radiation temperature, ai r temperature and air velocity
equivalent temperature；Dufton，1932
for heated room， 8C<tep<24C，V<0.5m/s
in1929，Dufton developed an integrated th ermostat：eupatheostat
Based on thermal equilibrium
predicted mean vote-predicted percent of d issatisfied （PMV-PPD）；Fanger，1972（m ost comprehensive and stable）
The cold and hot sensation near thermal ne utrality is mainly predicted. the 6 factors are considered comprehensively
human activities
clothing thermal resistance
air temperature
mean radiation temperature
air humidity
air velocity
Comfort equation PMV: an indicator that pr edicts the thermal sensation generated by a ny given combination of environmental vari ables
Fanger divided the result value into seven le vels
heat stress index-HSI；Belding Hatch1955
21C<ET<60C,0.25m/s<v<10m/s
The purpose of this method is to program a single heat index with comprehensive envir onmental variables to quantitatively express the stress of thermal environment on the h uman body
new effctive temperature-ET*；Gagge,Stolw ijk,Nishi,1971
sitting and light clothing
standard effctive temprrature-SET；Gagge,S tolwijk,Nishi,1971
proper in the thomerl range wothout chill
According to the physiological response bet ween the body and the environment heat ex change coefficient / empirical indicators
Wind Chill Index-WCI，Siple
V<20m/s
in very chilled weather, the main two factors which effect the human thomerl loss is air v elocity and temperature
effctive temperature-ET；Houghton，Yaglo u，1932
1C<ET<43C,0.1<V<3.5m/s
disatisfied index-DI；UK Met office，1957
A sultry environment is evaluated by a com bination of temperature and humidity

Figure 2. 5 Main Evaluation Indices of Thermal Comfort.

- PMV-PPD method

The PMV/PPD model was developed by P.O. Fanger （1972）to define comfort using thermal equilibrium equations and empirical research on skin temperature. The standard thermal comfort survey requires subjects to be asked how they feel about heat on a seven-point scale from cold (-3) to hot (+3). The Fanger equation is used to calculate the predicted mean value (PMV) for a group of subjects under a specific combination of air temperature, mean radiated temperature, relative humidity, relative humidity, air velocity, metabolic rate, and clothing insulation. PMV equals 0 for thermal neutrality, and the comfort zone is defined by a combination of six parameters with PMV within the recommended range (-0.5<PMV<+0.5). While predicting how hot the population will feel is an important step in determining what conditions are comfortable, considering whether people will be satisfied, Fanger developed another equation that relates PMV to the predicted percentage of dissatisfaction (PPD). This relationship is based on studies in which subjects are investigated under indoor conditions that can be precisely controlled. The PMV/PPD model is applied globally but does not directly consider adaptation mechanisms and outdoor thermal conditions. ASHRAE Standard 55-2017 uses the PMV model to set requirements for indoor thermal conditions. CBE's ASHRAE 55 Thermal Comfort Tool allows users to enter six comfort parameters to determine if a combination meets ASHRAE 55 standards. The results are shown on a psychometric or temperature-humidity graph and show that for a given temperature and relative humidity range, the input values of the remaining 4 parameters will make the temperature and relative humidity comfortable.

The PMV/PPD model has low predictive accuracy. Using the world's largest thermal comfort field survey database, PMV was only 34% accurate in predicting occupant thermal sensation, which is one-third correct. Whereas PPD outside the thermal intermediate range (-1≤PMV≤1) overestimated the subjects' thermal unacceptability. The accuracy of PMV/PPD varies considerably with different ventilation strategies, building types and climatic conditions. Human factors can have a significant impact on thermal comfort. However, the artificial factor cannot be generalized. This can be a big problem.

- Standard effective temperature

Standard effective temperature (SET) is a model of human response to the thermal environment. Developed by A.P. Gagge and accepted by ASHRAE in 1986, it is also referred to as the Pierce Two-Node model. Its calculation is similar to PMV because it is a comprehensive comfort index based on heat-balance equations that incorporates the personal factors of clothing and metabolic rate. Its fundamental difference is it takes a two-node method to represent human physiology in measuring skin temperature and skin wetness.

ASHRAE 55-2010 defines SET as "the temperature of an imaginary environment at 50% relative humidity, <0.1 m/s [0.33 ft/s] average air speed, and mean radiant temperature equal to average air temperature, in which total heat loss from the skin of an imaginary occupant with an activity level of 1.0 met and a clothing level of 0.6 clo is the same as that from a person in the actual environment, with actual clothing and activity level".

- Cooling effect

ASHRAE 55-2017 defines the Cooling Effect (CE) at elevated air speed (above 0.2 metres per second (0.66 ft/s)) as the value that, when subtracted from both the air temperature and the mean radiant temperature, yields the same SET value under still air (0.1 m/s) as in the first SET calculation under elevated air speed.

$$SET(t_a, t_r, v, met, clo, RH) = SET(t_a - CE, t_r - CE, v = 0.1, met, clo, RH)$$

The CE can be used to determine the PMV adjusted for an environment with elevated air speed using the adjusted temperature, the adjusted radiant temperature and still air (0.2 metres per second (0.66 ft/s)). Where the adjusted temperatures are equal to the original air and mean radiant temperatures minus the CE.

- Radiant Temperature Asymmetry

Significant differences in thermal radiation from the surfaces surrounding an individual can cause local discomfort or reduce the acceptance of thermal conditions. ASHRAE Standard 55 establishes limits on allowable temperature differences between various surfaces. Sensitivity to asymmetry varies depending on the surfaces involved; for example, people are more sensitive to a warm ceiling compared to hot and cold vertical surfaces. Therefore, the limits depend on which surfaces are affected. The ceiling temperature cannot exceed other surfaces by more than +5 °C (9.0 °F), while a wall may be up to +23 °C (41 °F) warmer than other surfaces.

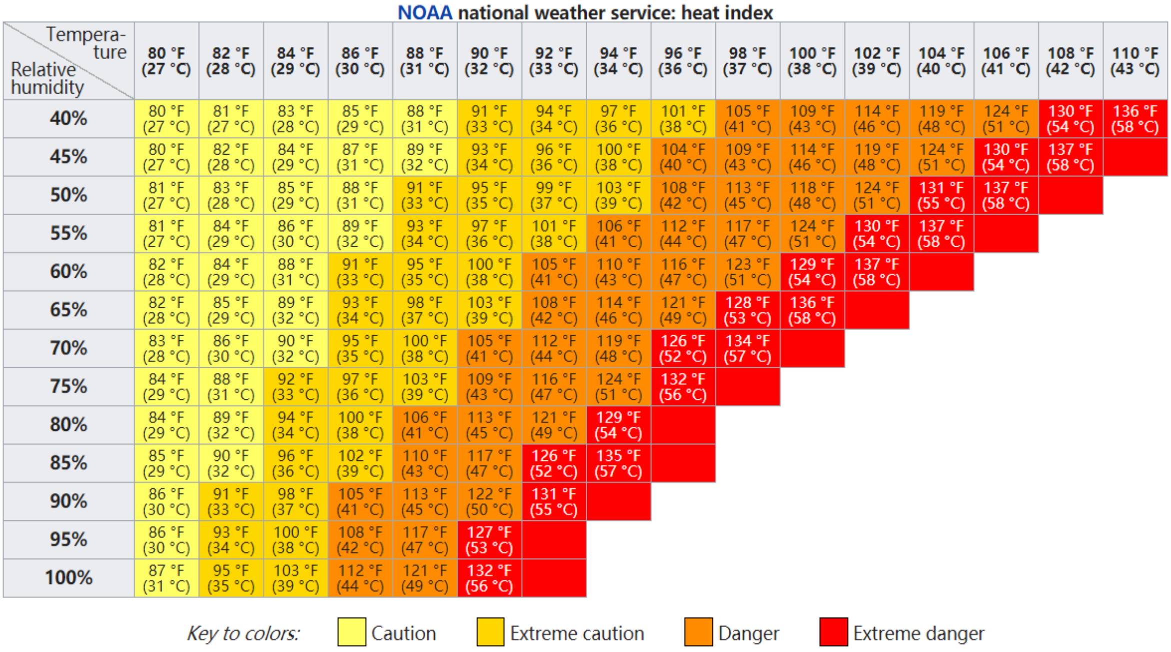
- Interaction of Temperature and Humidity

Both the psychrometric chart and the bioclimatic chart are based on this concept. Various indices, such as the heat index, have been developed to combine air temperature and humidity for higher temperatures. For lower temperatures, these interactions are only qualitatively understood rather than quantitatively measured. High humidity and low temperatures can create a chilling effect; cold air with high relative humidity "feels" colder than dry air of the same temperature because moisture increases the body’s heat transfer. The reason for this phenomenon is thought to be that high humidity makes skin and clothing moist, enhancing their conductivity and thereby increasing heat loss through conduction.

- Heat Index

The heat index (HI), also known as humiture, is an index that combines air temperature and relative humidity in a shaded area to provide an equivalent temperature perceived by the human body—essentially, how hot it would feel if the humidity level were different. The human body cools itself through sweating, with heat being expelled as sweat evaporates. However, high relative humidity reduces the rate of evaporation, resulting in slower heat removal from the body and a sensation of overheating. This effect is subjective, as individuals perceive heat differently due to various factors, such as body size, metabolism, hydration, pregnancy, menopause, drug effects, or withdrawal symptoms. The heat index calculation is based on temperatures measured in the shade. However, when people are active in direct sunlight, the actual felt temperature may be much higher than indicated by the heat index. For example, while a temperature of 28 °C (82 °F) in the shade with a relative humidity of 60% yields a heat index of 29 °C (84 °F), engaging in physical activity in full sunlight could raise the perceived heat index significantly—sometimes exceeding 58 °C (136 °F). This highlights that the heat index may underestimate the temperature unless the person is resting in a shaded area. The following Table 2.6 is from the National Oceanic and Atmospheric Administration of the United States（Weinberger et al., 2018）. These columns start at 80°F (27°C) but also have a heat index effect at 79°F (26°C) and high humidity.

Table 2. 6 National Weather Service: heat index



NOAA national weather service: heat index

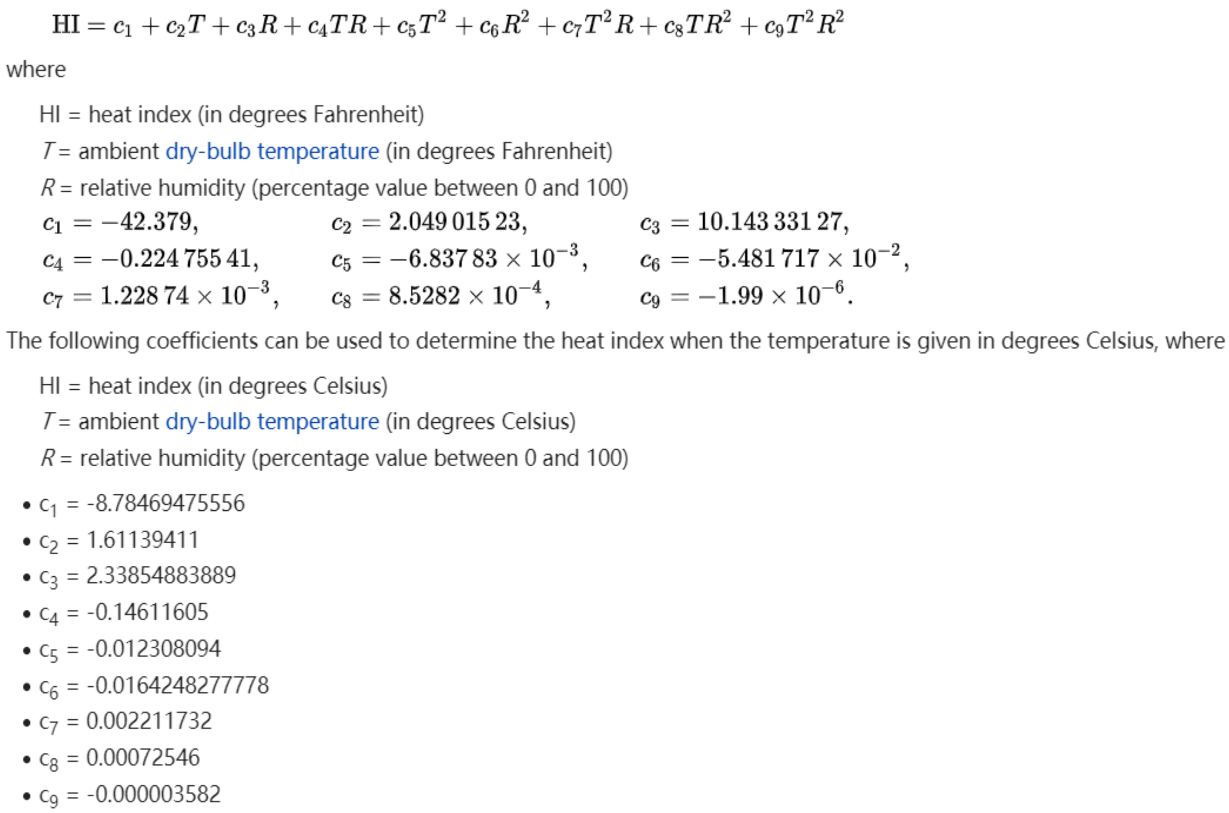
Temperature Relative humidity	80 °F (27 °C)	82 °F (28 °C)	84 °F (29 °C)	86 °F (30 °C)	88 °F (31 °C)	90 °F (32 °C)	92 °F (33 °C)	94 °F (34 °C)	96 °F (36 °C)	98 °F (37 °C)	100 °F (38 °C)	102 °F (39 °C)	104 °F (40 °C)	106 °F (41 °C)	108 °F (42 °C)	110 °F (43 °C)
40%	80 °F (27 °C)	81 °F (27 °C)	83 °F (28 °C)	85 °F (29 °C)	88 °F (31 °C)	91 °F (33 °C)	94 °F (34 °C)	97 °F (36 °C)	101 °F (38 °C)	105 °F (41 °C)	109 °F (43 °C)	114 °F (46 °C)	119 °F (48 °C)	124 °F (51 °C)	130 °F (54 °C)	136 °F (58 °C)
45%	80 °F (27 °C)	82 °F (28 °C)	84 °F (29 °C)	87 °F (31 °C)	89 °F (32 °C)	93 °F (34 °C)	96 °F (36 °C)	100 °F (38 °C)	104 °F (40 °C)	109 °F (43 °C)	114 °F (46 °C)	119 °F (48 °C)	124 °F (51 °C)	130 °F (54 °C)	137 °F (58 °C)	
50%	81 °F (27 °C)	83 °F (28 °C)	85 °F (29 °C)	88 °F (31 °C)	91 °F (33 °C)	95 °F (35 °C)	99 °F (37 °C)	103 °F (39 °C)	108 °F (42 °C)	113 °F (45 °C)	118 °F (48 °C)	124 °F (51 °C)	131 °F (55 °C)	137 °F (58 °C)		
55%	81 °F (27 °C)	84 °F (29 °C)	86 °F (30 °C)	89 °F (32 °C)	93 °F (34 °C)	97 °F (36 °C)	101 °F (38 °C)	106 °F (41 °C)	112 °F (44 °C)	117 °F (47 °C)	124 °F (51 °C)	130 °F (54 °C)	137 °F (58 °C)			
60%	82 °F (28 °C)	84 °F (29 °C)	88 °F (31 °C)	91 °F (33 °C)	95 °F (35 °C)	100 °F (38 °C)	105 °F (41 °C)	110 °F (43 °C)	116 °F (47 °C)	123 °F (51 °C)	129 °F (54 °C)	137 °F (58 °C)				
65%	82 °F (28 °C)	85 °F (29 °C)	89 °F (32 °C)	93 °F (34 °C)	98 °F (37 °C)	103 °F (39 °C)	108 °F (42 °C)	114 °F (46 °C)	121 °F (49 °C)	128 °F (53 °C)	136 °F (58 °C)					
70%	83 °F (28 °C)	86 °F (30 °C)	90 °F (32 °C)	95 °F (35 °C)	100 °F (38 °C)	105 °F (41 °C)	112 °F (44 °C)	119 °F (48 °C)	126 °F (52 °C)	134 °F (57 °C)						
75%	84 °F (29 °C)	88 °F (31 °C)	92 °F (33 °C)	97 °F (36 °C)	103 °F (39 °C)	109 °F (43 °C)	116 °F (47 °C)	124 °F (51 °C)	132 °F (56 °C)							
80%	84 °F (29 °C)	89 °F (32 °C)	94 °F (34 °C)	100 °F (38 °C)	106 °F (41 °C)	113 °F (45 °C)	121 °F (49 °C)	129 °F (54 °C)								
85%	85 °F (29 °C)	90 °F (32 °C)	96 °F (36 °C)	102 °F (39 °C)	110 °F (43 °C)	117 °F (47 °C)	126 °F (52 °C)	135 °F (57 °C)								
90%	86 °F (30 °C)	91 °F (33 °C)	98 °F (37 °C)	105 °F (41 °C)	113 °F (45 °C)	122 °F (50 °C)	131 °F (55 °C)									
95%	86 °F (30 °C)	93 °F (34 °C)	100 °F (38 °C)	108 °F (42 °C)	117 °F (47 °C)	127 °F (53 °C)										
100%	87 °F (31 °C)	95 °F (35 °C)	103 °F (39 °C)	112 °F (44 °C)	121 °F (49 °C)	132 °F (56 °C)										

Key to colors: □ Caution □ Extreme caution □ Danger □ Extreme danger

HI formula

Several HI formulas have been developed by Anderson et al. (2013), NWS (2011), Jonson and Long (2004), and Schoen (2005). The first two are polynomial based, while the third uses a single formula involving exponential functions.

The following formula approximates the heat index (HI) in degrees Fahrenheit with an accuracy of ±1.3 °F (0.7 °C). It results from a multivariate fit based on human body modelling for temperatures equal to or greater than 80 °F (27 °C) and relative humidity equal to or greater than 40%. This equation closely matches the NOAA National Weather Service table, with minor deviations (less than ±1, respectively) at 90 °F (32 °C) and 45%/70% relative humidity when unrounded.



$$\mathrm{HI} = c_1 + c_2T + c_3R + c_4TR + c_5T^2 + c_6R^2 + c_7T^2R + c_8TR^2 + c_9T^2R^2$$

where

HI = heat index (in degrees Fahrenheit)
T = ambient dry-bulb temperature (in degrees Fahrenheit)
R = relative humidity (percentage value between 0 and 100)

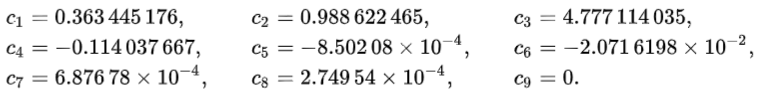
$$\begin{aligned} c_1 &= -42.379, & c_2 &= 2.049\,015\,23, & c_3 &= 10.143\,331\,27, \\ c_4 &= -0.224\,755\,41, & c_5 &= -6.837\,83 \times 10^{-3}, & c_6 &= -5.481\,717 \times 10^{-2}, \\ c_7 &= 1.228\,74 \times 10^{-3}, & c_8 &= 8.5282 \times 10^{-4}, & c_9 &= -1.99 \times 10^{-6}. \end{aligned}$$

The following coefficients can be used to determine the heat index when the temperature is given in degrees Celsius, where

HI = heat index (in degrees Celsius)
T = ambient dry-bulb temperature (in degrees Celsius)
R = relative humidity (percentage value between 0 and 100)

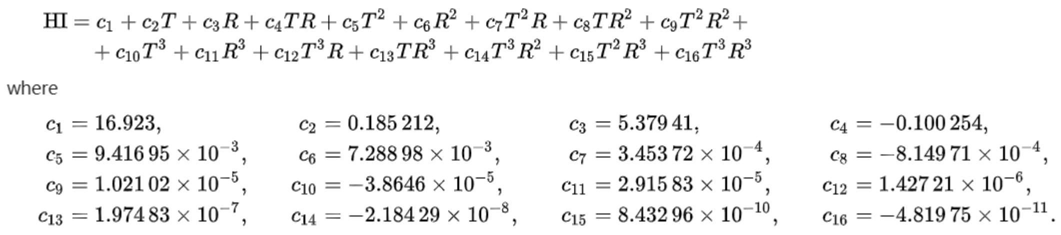
- c_1 = -8.78469475556
- c_2 = 1.61139411
- c_3 = 2.33854883889
- c_4 = -0.14611605
- c_5 = -0.012308094
- c_6 = -0.0164248277778
- c_7 = 0.002211732
- c_8 = 0.00072546
- c_9 = -0.000003582

An alternative set of constants for this equation that is within ±3 °F (1.7 °C) of the NWS master table for all humidities from 0 to 80% and all temperatures between 70 and 115 °F (21–46 °C) and all heat indices below 150 °F (66 °C) is:



$$\begin{aligned} c_1 &= 0.363\,445\,176, & c_2 &= 0.988\,622\,465, & c_3 &= 4.777\,114\,035, \\ c_4 &= -0.114\,037\,667, & c_5 &= -8.502\,08 \times 10^{-4}, & c_6 &= -2.071\,6198 \times 10^{-2}, \\ c_7 &= 6.876\,78 \times 10^{-4}, & c_8 &= 2.749\,54 \times 10^{-4}, & c_9 &= 0. \end{aligned}$$

A further alternate is this:



$$\begin{aligned} \mathrm{HI} = {} & c_1 + c_2T + c_3R + c_4TR + c_5T^2 + c_6R^2 + c_7T^2R + c_8TR^2 + c_9T^2R^2 + \\ & + c_{10}T^3 + c_{11}R^3 + c_{12}T^3R + c_{13}TR^3 + c_{14}T^3R^2 + c_{15}T^2R^3 + c_{16}T^3R^3 \end{aligned}$$

where

$$\begin{aligned} c_1 &= 16.923, & c_2 &= 0.185\,212, & c_3 &= 5.379\,41, & c_4 &= -0.100\,254, \\ c_5 &= 9.416\,95 \times 10^{-3}, & c_6 &= 7.288\,98 \times 10^{-3}, & c_7 &= 3.453\,72 \times 10^{-4}, & c_8 &= -8.149\,71 \times 10^{-4}, \\ c_9 &= 1.021\,02 \times 10^{-5}, & c_{10} &= -3.8646 \times 10^{-5}, & c_{11} &= 2.915\,83 \times 10^{-5}, & c_{12} &= 1.427\,21 \times 10^{-6}, \\ c_{13} &= 1.974\,83 \times 10^{-7}, & c_{14} &= -2.184\,29 \times 10^{-8}, & c_{15} &= 8.432\,96 \times 10^{-10}, & c_{16} &= -4.819\,75 \times 10^{-11}. \end{aligned}$$

For example, using the last formula with a temperature of 90°F (32°C) and a relative humidity of 85% results in a thermal index of 90°F and a relative humidity of 85% = 114.9. This formula can be used directly as a knowledge base for HI, and can be used directly for HI, although the HI is only for shade conditions, but it is also possible for the environment inside the building.

Adaptive Comfort Mode

The adaptive model can also be applied in outdoor environments, incorporating more natural elements and offering several advantages over the PMV-PDD model. This model is based on the idea that outdoor climate influences indoor comfort, as people adapt to different temperatures throughout the year. The adaptive hypothesis suggests that environmental factors, such as access to environmental controls and an individual's thermal history, affect the thermal expectations and preferences of building occupants. Numerous researchers worldwide have conducted field studies investigating the thermal comfort of building users alongside environmental measurements. An analysis of data from 160 buildings revealed that occupants of naturally ventilated buildings accept, and even prefer, a wider range of temperatures than those in sealed, air-conditioned buildings, as their temperature preferences are influenced by outdoor conditions. These findings were incorporated into the ASHRAE 55-2004 standard as the adaptive comfort model. The adaptive model graph correlates indoor comfort temperature with the prevailing outdoor temperature and defines the 80% and 90% satisfaction zones. The model is particularly suited for naturally ventilated spaces where occupants have control over the environment, and where the outdoor climate influences indoor conditions and, consequently, the comfort zone. Research by de Dear and Brager has shown that occupants in naturally ventilated buildings exhibit greater tolerance for temperature variations due to behavioural and physiological adjustments, as different types of adaptive processes are at play. ASHRAE Standard 55-2010 states that recent thermal experience, clothing adjustments, access to control options, and changes in occupant expectations can all influence thermal responses.

Adaptive models for thermal comfort are also implemented in other standards, such as the European EN 15251 and ISO 7730 standards. Although their derivation methods and results differ slightly from the ASHRAE 55 adaptive standard, they are fundamentally similar. The main difference lies in their applicability: the ASHRAE adaptive standard applies only to buildings without mechanical cooling, while the EN 15251 standard can also be applied to mixed-mode buildings, provided the mechanical system is not in operation.

Thermal adaptations are broadly classified into three categories: behavioral adaptations, physiological adaptations, and psychological adaptations. This model is much more complex than the previous one, as behavioural, physical and psychological adaptations need to be taken into account.

Regional differences

In different parts of the world, the need for thermal comfort varies due to different climates. China's climate is hot and humid in the summer and cold in the winter, which creates a need for thermal comfort.

The relationship between energy efficiency and thermal comfort has become a big issue in China over the past few decades due to rapid economic and population growth. Currently, researchers are looking at ways to heat and cool buildings in China at a lower cost and less harmful to the environment. In the Brazilian tropics, urbanization is creating urban heat islands (UHI). These refer to urban areas that have exceeded the thermal comfort limit due to the influx of people and only fall into the comfort range during the rainy season. Urban heat islands can occur over any urban city or built-up area under the right conditions. In this hot and humid region of Saudi Arabia, the issue of thermal comfort has always been an important one in the mosques where Muslims go to pray. They are very large open buildings that are only used intermittently (very busy at noon Friday prayers) and are difficult to ventilate properly. The size of the building requires a lot of ventilation, but this requires a lot of energy, as the buildings are constructed from mosques.

Use only for a short period of time. Some mosques run their HVAC systems too long, and some mosques stay too cold. Due to the large size of the mosque, the stacking effect also occurs, creating a large layer of hot air above the people in the mosque. The new design places the ventilation system lower in the building to provide more temperature control at ground level. In addition, new controls have been put in place to improve efficiency.

The impact of the Reginal variance on the building is important. This affects all natural factors, so people also produce different human factors.

2.4.3 Bioclimate chart and Psychrometric chart

The bioclimatic chart is the most recognised quantitative chart to define the comfort zone, which integrates environmental and physiological impacts. And the bioclimatic chart developed gradually after 1963 and can now be applied to the design of HVAC in buildings.

The Bioclimatic Chart was developed initially by Victor Olgyay and used in the case of passive cooling methods (i.e., without mechanical assistance). After that, it was developed by Barruch Givoni in 1976, modified by Murray Milne in 1979, Watson and Labs again further modified and enhanced the building bioclimatic Chart or Psychrometric Chart in 1983. Finally, ASHRAE developed the comfort chart. The latter two systems are considered the mechanical approach and are now commonly used in the design of HVAC systems.

In Olgyay's bioclimatic chart, dry bulb temperature (DBT), relative humidity, mean radiant temperature, wind speed, and solar radiation are combined as different thermal factors to define a comfort zone. The surrounding climatic elements are represented using curves that indicate the type of corrective measures needed to restore comfort for any point outside the comfort zone. In 1976, Barruch Givoni proposed the Building Bioclimatic Chart based on the ASHRAE humid air diagram, incorporating the building as a factor. This chart illustrates a linear relationship between temperature amplitude and vapor pressure in the outdoor environment (Figure 2.6). Compared to Olgyay's bioclimatic chart, Givoni's model accounts for the presence of buildings and removes geographical limitations (Al-Azri et al., 2013). Both charts define the same boundaries. The acceptable temperature range is approximately 21°C to 27.5°C, and the comfort zone also specifies a relative humidity range of 20% to 78%.

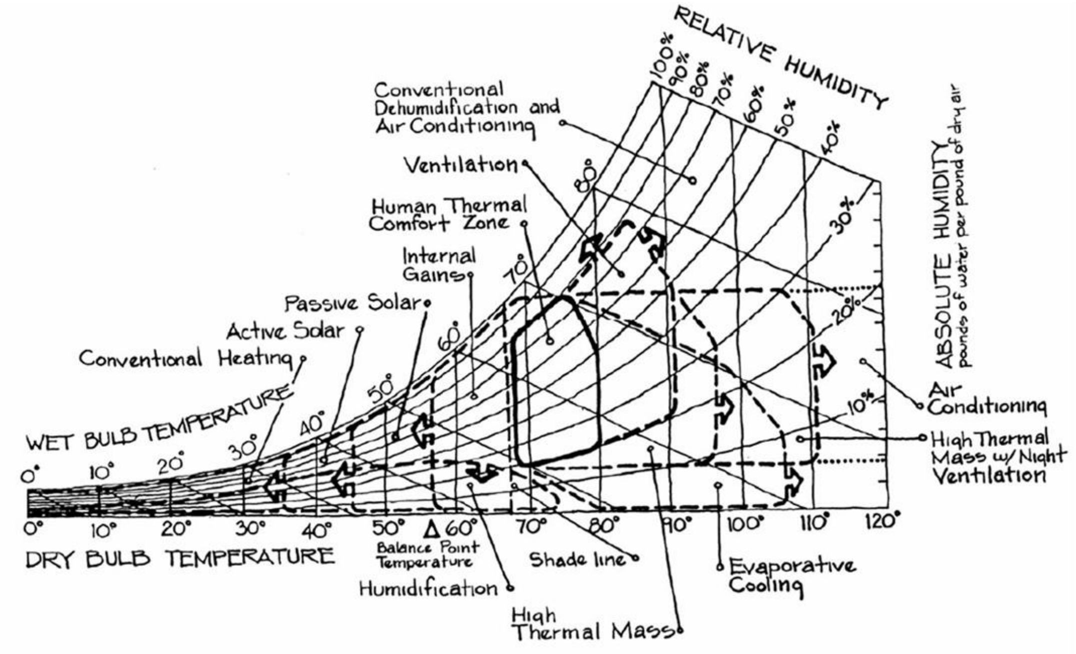


Figure 2. 6 Building Bio-climatic Chart (Barruch, 1976).

A psychometric chart is a graphical representation of the psychometric process of air. Psychometric processes include physical and thermodynamic properties such as dry ball temperature, wet ball temperature, humidity, enthalpy and air density. Psychometric mapping can be done in two different ways. The first is done by plotting multiple data points on a graph that represent air conditions at a given time. Then, cover an area to determine the "comfort zone". A comfort zone is defined as the extent to which the occupants are satisfied with the thermal conditions around them. After mapping out the air conditions and superimposing the comfort zone, one can see how the passive design strategy extends the comfort zone. The diagram is also often used by mechanical engineers to dynamically plot points that represent outside air conditions and to understand the processes that air must go through to achieve conditions that are comfortable for occupants inside the building. When a psychometric chart is used for this purpose, the data points move across the chart.

According to the bioclimatic charts above and the psychrometric chart, the factors which influence the thermal comfort in a building could be concluded:

- Air temperature
- Radiant temperature
- Relative humidity
- Air movement/ velocity

According to Loong's note in City University of Hong Kong, Dry bulb temperature does not consider air humidity and radiation. It can be changed by cooling or heating, or by introducing fresh air of a different temperature. Changing the relative humidity can be influenced by injecting water droplets or steam into the air, or allowing air to pass through the evaporation surface. Besides, passing the air over a (cooling and) de-humidifying coil, a chemical absorber, or passing the air over an air washer supplied with chilled water could also affect the relative humidity.

Referring to ANSI/ASHRAE Standard 55(2010), the comfort zone the combinations of air temperature, mean radiant temperature (tr), and humidity that are predicted to be an acceptable thermal environment at particular values of airspeed, metabolic rate, and clothing insulation (Figure 2.7).

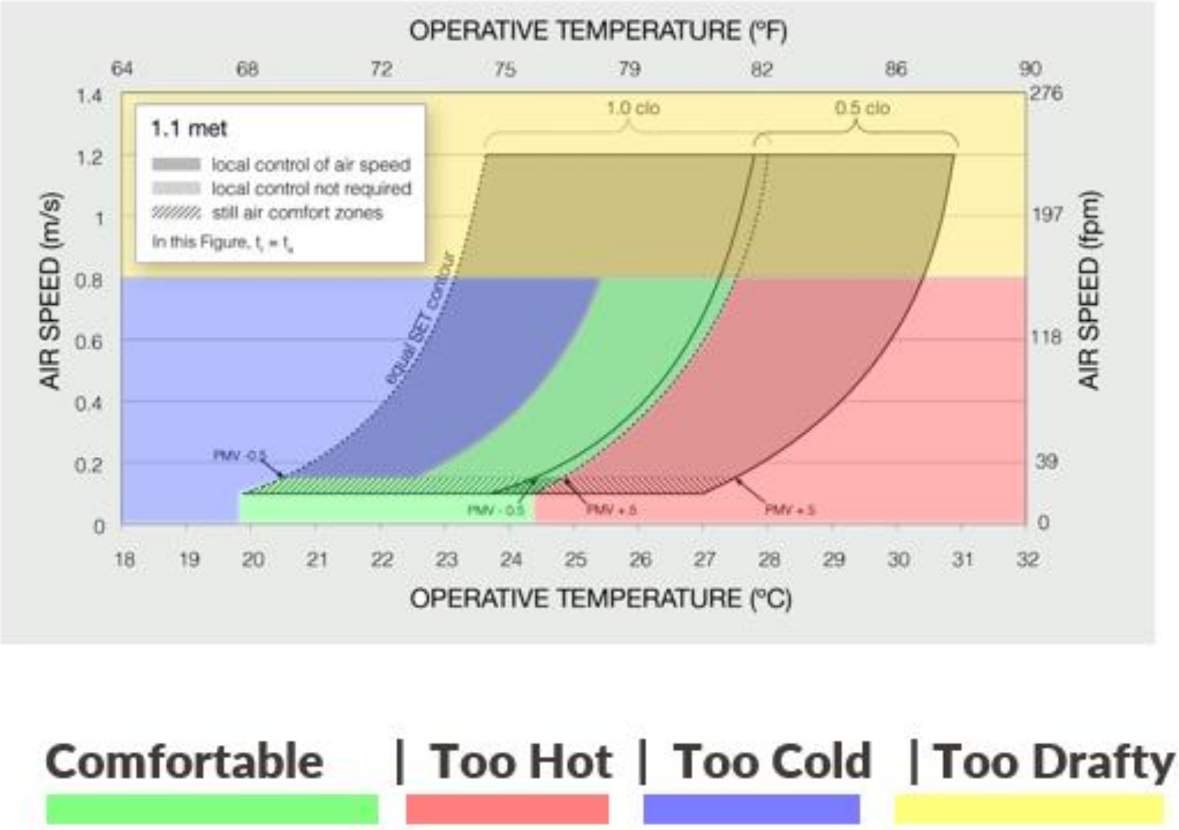


Figure 2. 7 Thermal comfort chart using ASHRAE 55 parameters

Typically, the attributes represented by a psychometric chart are:

- Dry Ball Temperature: A measurement of air temperature recorded with a thermometer, exposed to air but unaffected by radiation and moisture.
- Wet ball temperature: a thermometer recorder with a cloth wrapped around the ball, wetting the ball with distilled water. A wet ball evaporates at different rates, so it records different temperatures depending on the humidity of the air it is exposed to.
- Relative humidity: the ratio of the actual vapour pressure to the vapour pressure of saturated air at the same temperature, expressed as a percentage.
- Specific volume: The volume of dry air per weight.
- Dew point temperature: The maximum temperature at which water vapour condenses.
- Humidity Ratio: The dry base moisture content of air, expressed as the weight of water vapour in a unit weight of dry air.
- Enthalpy: The energy content of air.

The condition of the moist air can be determined from any two of these properties and then all the others. Atmospheric pressure varies with altitude, so there are many psychometric maps available for different atmospheric pressures, but for sea level psychometric maps below 600 m above sea level are generally considered adequate.

Psychometric maps are more general and contain more variables than bioclimatic maps.

Common Applications While psychometric principles apply to any physical system consisting of a gas-vapor mixture, the most common system is a mixture of water vapor and air because of its application in heating, ventilation, air conditioning, and meteorology. As far as humans are concerned, our thermal comfort depends to a large extent not only on the temperature of the surrounding air, but also (because we cool down by sweating) on how saturated the air is with water vapor. Many substances are hygroscopic, which means they attract water, often in proportion to or above critical relative humidity. These include cotton, paper, cellulose, other wood products, sugar, calcium oxide (burned lime) and many chemicals and fertilizers. Industries that use these substances pay attention to relative humidity control in both their production and storage. In industrial drying applications, such as paper drying, manufacturers often try to achieve an optimum between low relative humidity and energy use, where low relative humidity increases the drying rate and energy use decreases as the exhaust relative humidity increases. In many industrial applications, it is important to avoid condensation, which can damage products or cause corrosion. Molds and fungi can be controlled by keeping the relative humidity low. Wood-destroying fungi generally do not grow at relative humidity below 75%.

2.4.4 Visual Comfort: Lighting and its Effects

Light is a big part of energy consumption. It accounts for 1/3 of total energy consumption in shopping and office buildings. hence, it is of great social significance and economic benefit to reduce the energy consumption of lighting to the greatest extent under the condition that the indoor light environment is satisfied.

2.4.4.1 Basic Theories

Light is radiant energy that travels in the form of electromagnetic waves:

- Visible light： Wavelength is in the range of 380nm~760nm
    - different invisible light has different colours; Monochromatic light: light of a single wavelength
    - Daylight and light are composite: white or another colour.
    - The spectral power distribution curve of composite light is formed by linking the relative power quantities of radiation at various wavelengths in the composite light according to the corresponding wavelength arrangement.
- Wavelength<380nm: Ultraviolet (uv); X-rays, Y-rays, cosmic rays
- Wavelength>760nm: Infrared ray; Radio waves

There are four common light metrics:

- Luminous flux（1umen/1m）
- Intensity of illumination（1ux, 1x）
- Luminous intensity(candela/cd)
- Luminance（nit/nt）1nt=cd/m2

Physical brightness: it is distinguished from the visual perception of light and darkness. As the eye adapts to the brightness of the environment, the visual perception of an object may be higher or lower than its physical brightness.

2.4.4.2 Indoor light environment

In optimizing the indoor light environment, particularly daylight, it's crucial to capitalize on sunlight's positive aspects while ensuring living spaces receive a sufficient quantity during the colder months. Winter sunlight regulations should specify the minimum duration and intensity of sunlight exposure in living areas to foster a comfortable and healthy environment. This minimum is typically dictated by the sun's lowest winter trajectory and the need to mitigate indoor pathogens and bolster health, reflecting local climatic conditions and health considerations. Conversely, mitigating sunlight during the hotter months is equally important to prevent excessive indoor heat.

Several factors dictate the quality and quantity of indoor sunlight, including:

- The orientation of the space, determining the sun's path relative to the room.
- The building's exposure structure, influences how sunlight enters and moves within the space.
- Local photo climatic conditions that vary with geography and climate.
- The specific geographical conditions of the site.

Regarding artificial lighting, it's imperative to strike a balance that mimics natural light as closely as possible. The window-to-floor area ratio, angles of light entry, and the naturalness factor of lighting all contribute to a well-lit and comfortable space. Sanitary requirements dictate that indoor lighting should provide stable, uniform illumination without causing glare or discomfort, with a spectral composition akin to natural daylight.

Light comfort pivots on four key criteria:

- Adequate illumination levels that cater to both the task at hand and overall comfort.
- A harmonious brightness ratio that avoids abrupt contrasts and reduces glare.
- An optimal colour temperature and rendering that fosters an accurate and pleasant visual experience.
- Diligent avoidance of glare and disruptive light sources.

Light pollution emerges as a significant concern in modern environments, predominantly caused by excessive or improperly directed artificial light. It's essential to mitigate such pollution by employing thoughtful design and technology, ensuring a harmonious blend of artificial and natural light to sustain healthful and pleasant indoor atmospheres.

2.4.5 Acoustic Comfort

2.4.5.1 Noise and its Impact

There are three primary elements of sound generation and transmission: sources, pathways, and recipients. The acoustic source is a vibrating object that produces sound waves, which travel through a medium（Vardaxis，2018）. This medium often includes the outdoor atmosphere, indoor air, walls, floorboards, and other structural components. The recipient of the sound is the human ear, and sound is typically characterized by three properties: volume, pitch, and tone.

A sound wave is a type of pressure wave, essentially a fluctuation of air pressure around a static pressure level. The variation in air pressure caused by the sound wave is referred to as sound pressure. The propagation of sound involves the transmission of pressure waves, and while air particles oscillate back and forth around their equilibrium positions, they do not travel with the wave itself. The speed at which these pressure waves propagate is the speed of sound, whereas the speed of oscillation of air particles corresponds to the intensity or strength of the sound. Noise is generally defined as any unwanted or disruptive sound. It can vary widely in its characteristics and impact, which necessitates a range of evaluation methods to accurately assess its effects.

Noise evaluation encompasses numerous methods, considering several influencing factors: Firstly, noise intensity, spectral characteristics, and temporal aspects such as onset time, duration, and fluctuations are critical in understanding how noise affects individuals. For instance, a sudden, loud noise might be more disruptive than a continuous, low-level hum. The time of day when noise occurs and its duration also play a role in determining its impact on comfort and well-being. Secondly, the context in which noise occurs, including the nature of people's lifestyles, work environments, and overall environmental conditions, significantly influences how noise is perceived and its subsequent effects. In environments where concentration is crucial, such as offices or educational institutions, even moderate noise levels can be more disruptive than in a more relaxed setting like a park. Thirdly, individual auditory properties also influence how noise affects a person, relating to both physiological and psychological responses. For example, the same sound might cause discomfort or stress in one person but might be perceived as negligible by another, highlighting the importance of considering individual sensitivities and health conditions when evaluating noise.

Additionally, the measurement conditions and methods used in noise evaluation must be standardized to ensure consistency. The conditions under which noise is measured, such as the type of equipment, measurement distance, and environmental factors (e.g., temperature, wind), can all influence the accuracy and reliability of the evaluation. Standardization ensures that noise measurements are comparable across different locations and periods.

2.4.5.2 Common Noise Evaluation Methods and Indicators:

A-Weighted Sound Level (dBA):

This is the most commonly used method worldwide for assessing noise levels. It measures noise in a way that reflects the human ear's sensitivity, making it particularly useful for steady-state noise like background traffic or machinery. However, it has limitations, as it does not provide detailed information about the noise spectrum. For instance, it might not differentiate between high-pitched or low-pitched noises, which could be relevant depending on the noise source.

Equivalent Continuous A-Weighted Sound Level (Leq):

The Leq method is designed for situations where noise levels fluctuate over time. By measuring noise at intervals and averaging these readings, Leq provides a comprehensive picture of average noise exposure over a period. However, it is less effective in capturing the impact of occasional, short bursts of noise, such as sudden loud events, which may still cause discomfort despite their brevity.

Day-Night Equivalent Sound Level (Ldn):

The Ldn method accounts for variations in noise sensitivity during different times of the day by adding a 10 dB penalty to nighttime noise levels. This reflects the increased sensitivity and potential disturbances caused during nighttime when people are more likely to seek rest or quiet. It’s particularly useful for urban planning and noise regulation, ensuring that residential areas remain conducive to sleep and rest.

Cumulative Distribution of Sound Levels:

This statistical approach measures noise by determining the percentage of time a sound level exceeds a particular threshold. By analysing the cumulative distribution of noise levels over time, this method provides insights into the variability and frequency of noise exposure. It is useful for identifying patterns and understanding the probability of noise levels exceeding acceptable limits in specific environments.

Noise Rating (NR) and Noise Criteria (NC), PNC Curves:

The NR and NC curves are developed standards used to evaluate indoor noise levels, particularly in environments where speech clarity and comfort are essential, such as offices, schools, and hospitals. The ISO standard includes the NR curve, while the NC curves (proposed by Beranek in 1957 and recommended by ISO in 1968) are more stringent, especially for low-frequency noise. The PNC (Preferred Noise Criteria), a modification of the NC curve, further refines these standards to provide a more precise assessment. These curves are essential for ensuring that indoor spaces maintain an environment conducive to concentration and communication.

Health Effects of Noise:

Noise exposure can lead to a range of health effects, impacting the auditory system, digestive system, nervous system, female physiology, and visual organs. For example, continuous exposure to high sound pressure and high-frequency noise can cause hearing loss, while intermittent exposure to non-stationary noise might result in stress and fatigue. Long-term exposure can also contribute to cardiovascular issues, sleep disturbances, and even negative effects on pregnancy. Research shows that high-intensity noise not only affects hearing but also triggers physiological stress responses that impact other bodily systems, highlighting the need for comprehensive noise management strategies in both workplaces and residential areas.

2.5 Intersections of BIM, Machine Learning, and Building Comfort

This section reviews published articles sourced from major academic databases, including ScienceDirect, Scopus, and Google Scholar, covering the period from 2010 to July 2019. The keywords used in the search included: "Building Information Modelling" and "Building Energy" and "Data"; "Machine Learning" and "Building Energy"; "Building Information Modelling" and "Machine Learning." The inclusion criteria for the review were as follows:

- Studies published in English or other languages with English abstracts.
- Only research articles were considered.
- All specified keywords must be included in the articles.

The exclusion criteria were:

- Articles that were uncorrelated to the main topic.
- Articles published before 2010, deemed outdated for this study.

The review results are presented across three major fields:

BIM Applied to Energy Simulation: Studies in this area focus on how Building Information Modelling (BIM) is utilized for building energy simulations, which are instrumental in predicting energy performance and identifying efficiency opportunities during the design phase.

Machine Learning (ML) for Building Energy Consumption: This category covers the application of machine learning techniques for predicting, optimizing, and managing energy consumption in buildings. ML models provide advanced data analysis and pattern recognition, significantly enhancing the accuracy of energy forecasts.

Machine Learning Integrated with BIM: This field explores the synergy between ML and BIM, highlighting studies that combine these technologies to enhance building energy analysis, performance predictions, and optimization strategies.

As illustrated in Figure 2.8, the number of studies focusing on ML for building energy consumption has increased significantly over the years, reflecting the global concern for energy efficiency and sustainability. This upward trend demonstrates the growing importance of leveraging ML technologies to address the complexities of energy management in buildings. While BIM-based building energy analysis also shows moderate growth, the number of studies in this area remains smaller compared to those focused on ML for building energy consumption. This indicates that, while BIM is gaining traction in the energy simulation domain, it is still less frequently explored than standalone ML approaches. The findings also reveal that research combining BIM and ML remains limited. This suggests a potential opportunity for further exploration and integration, as the combined use of these technologies could provide more comprehensive and efficient solutions for building energy management and optimization.

.

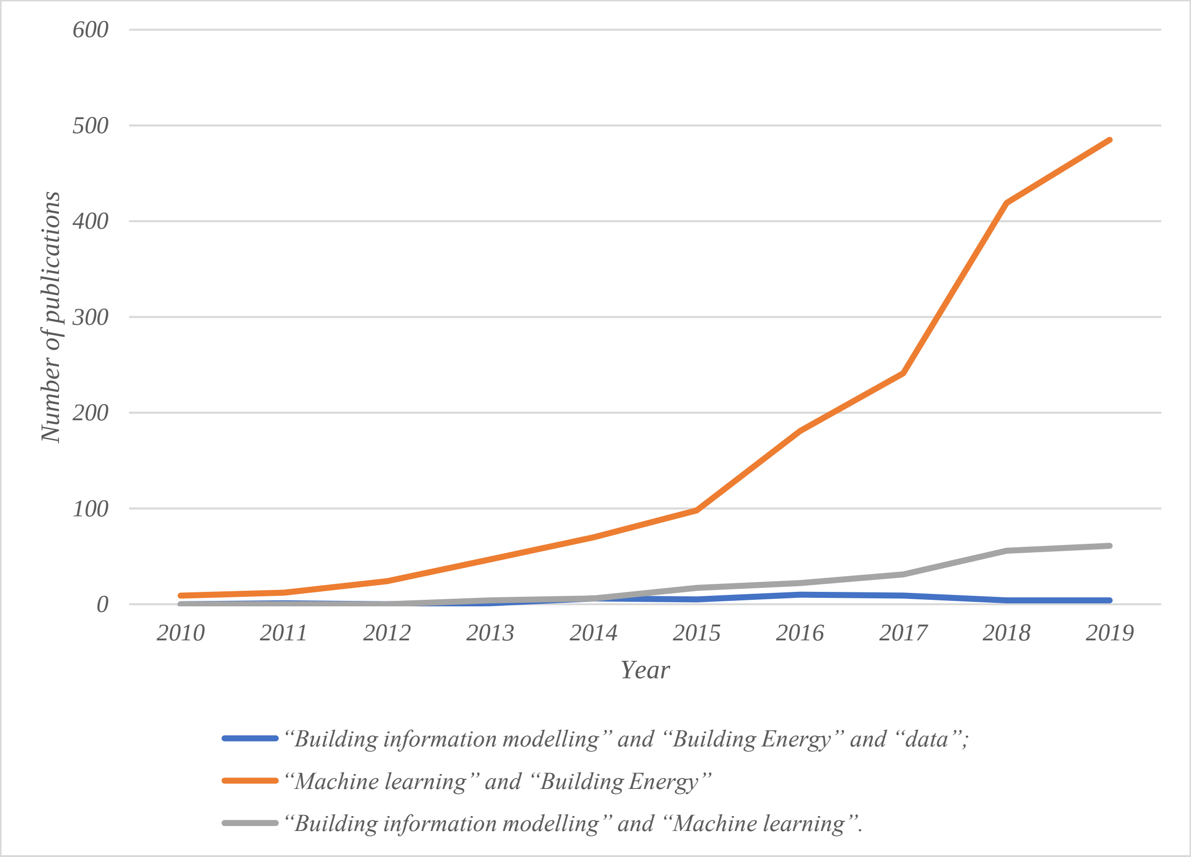


Figure 2. 8 The Quantity of publications in different domain

2.5.1 BIM supporting Building Energy Analysis

BIM is a method monitoring the whole life cycle of a construction (Eleftheriadis, Mumovic, & Greening, 2017). BIM captures multi-dimensional CAD information (Eadie, Browne, Odeyinka, McKeown, & McNiff, 2013). This digital revolution boosts the development of the AEC industry. The data can be easily accumulated and collected by the integration of BIM models. The concept of BIM is referred by the Krygiel et al. (2008) to an integrated database which stored all parametric and interconnected information of the entire building, and design documents. It will be reflected instantly throughout the rest of the project in all views if there are any changes to an object in the model. Additionally, as more systems have been added to buildings, the more energy is demanded to operate them.

There are many contributions of BIM in building energy domain, such as automation of energy modelling, enhancing the existing libraries, and storing and organizing the building data (Kamel & Memari, 2019). The main format of the BIM file is gbXML and IFC, which are generated from the BIM tool such as Revit. The research in 2015 developed a ModelicaBIM library for BIM-based building energy simulation (Kim, Jeong, Clayton, Haberl, & Yan, 2015). The ModelicaBIM library was used to investigate the system interface between BIM and energy simulation. This system interface can semi-automatically translate the building models in BIM to building energy models.

2.5.2 ML Supporting Energy Consumption Prediction

Machine learning (ML) methods have recently been applied in various domains, including the prediction of building energy consumption. With hundreds of studies in this area, ML techniques are becoming increasingly important in achieving the goals of energy efficiency and minimizing environmental impact. Accurate predictions of building energy consumption are crucial for optimizing building performance, improving energy management strategies, and reducing overall energy use. For instance, a review study conducted in 2012 categorized building energy prediction methods into four main types: engineering methods, statistical methods (such as regression), artificial intelligence (AI) methods, and the grey model (Zhao & Magoulès, 2012). The AI methods primarily include artificial neural networks (ANNs), support vector machines (SVMs), and decision trees, among others. In 2018, another review summarized energy consumption prediction models, emphasizing the importance of data properties, algorithms, and performance metrics (Amasyali & El-Gohary, 2018). This review provided a comprehensive overview of the different approaches and highlighted the key factors that influence the accuracy and efficiency of these models.

Among the AI methods, the artificial neural network (ANN) remains the most popular technique for building energy management (Bilal et al., 2016). ANNs have proven highly effective in solving non-linear and complex problems, making them suitable for modelling and predicting energy consumption in buildings. One of the critical aspects of using ANNs is the quality of input data. Raw data often contain noise, which can affect the performance of the model. To address this issue, researchers have developed various data pre-processing techniques to clean and normalize the data, ensuring that the models receive accurate and relevant information. In 2018, a study applied an ML approach to forecast building energy consumption (Dan & Phuc, 2018). The dataset used in this study consisted of historical energy usage data and building design parameters. By incorporating these variables, the researchers aimed to develop a more accurate and reliable model for predicting energy consumption, demonstrating the value of integrating building characteristics and historical data in ML models.

Overall, the use of ML in building energy prediction continues to grow, and as researchers refine these models and methods, the potential for further optimization of building energy performance becomes increasingly evident.

2.5.3 The BIM-based ML Development

Many machine learning (ML) methods have been fine-tuned and applied across various fields. Over the past decade, ML techniques have been used for project progress recognition on construction sites, utilizing site photos and BIM models (Golparvar-Fard, Peña-Mora, and Savarese, 2012). Subsequently, an integrated framework coordinating sensors and BIM elements was proposed by Bogen et al. (Bogen, Rashid, East, & Ross, 2013). This framework aimed to compare the operational state of building facilities with their scheduled state. K-means and hierarchical clustering methods were applied to classify resource usage based on typical human-specified schedules. In 2014, an experiment employed deep learning methods to classify 3D models within the BIM environment, yielding promising results (Qin, Li, Gao, Yang, & Chen, 2014). In 2015, research focused on the semi-automation and automation of collecting and processing photos from infrastructure construction sites (Teizer, 2015). This study utilized image recognition techniques to capture, analyse, and document the construction process. Building on previous studies, in 2016, research achieved successful 3D façade modelling and material recognition through photo-based recognition of as-built structures (Yang, Shi, & Wu, 2016). In 2017, ML was applied to basic clash detection in construction safety, enhancing safety protocols (Tixier, Hallowell, Rajagopalan, & Bowman, 2017). Simultaneously, a web-based platform was developed, based on the construction material library (CML) and BIM advancements, to simplify data collection and annotate material patches according to BIM overlays (Han & Golparvar-Fard, 2017). In 2018, a machine learning application for semantic enrichment of BIM models was proposed, focusing on extensive data pre-processing (Bloch & Sacks, 2018a). Another study that year demonstrated ML's direct applicability to space classification problems (Bloch & Sacks, 2018b). Similarly, ML techniques were utilized to distinguish relevant from irrelevant clashes, improving the quality of clash detection (Hu & Castro-Lacouture, 2018). In 2019, researchers validated that predictions generated by ML models were more representative of actual building performance compared to simulation results (Chen, 2019). Concurrently, a deep convolutional neural network was applied to indoor localization, successfully recognizing synthetic images from 3D indoor models (Acharya, Khoshelham, & Winter, 2019). Another study proposed an AI-based method to generate building designs according to client requirements, although the automation was limited to window design (Karan & Asadi, 2019). Additionally, natural language processing (NLP) and unsupervised learning were used to automatically identify whether a case study involved BIM (Jung & Lee, 2019).

Based on previous studies, the main applications and implications of BIM-based ML have been highlighted. This interdisciplinary domain requires further exploration to fully realize the potential of integrating BIM and ML technologies.

2.6 Fundamental Concepts of Ontology

2.6.1 What is Ontology?

Ontology has varied definitions across different disciplines, yet its foundational concepts remain similar. Primarily, it involves the study of the nature of being and categorization, as well as how entities in the real world are grouped and interpreted.

Philosophical Ontology: In the realm of philosophy, ontology is a branch of metaphysics that focuses on the study of the essence of existence (Uschold and Gruninger,1996). It involves the exploration of fundamental categories such as being, entities, processes, properties, space, and time. Philosophers seek to understand the intrinsic characteristics of various things in the world and how these elements constitute reality.

Information Science and Computer Science Ontology: In information science and computer science, ontology refers to an explicit specification of a domain's concepts and categories, along with the relationships between them. In this context, ontology is commonly used in data modeling, artificial intelligence, the Semantic Web, software engineering, and other fields to support the sharing and reuse of complex data. For instance, a healthcare ontology might define concepts like diseases, symptoms, treatments, and patients, and the relationships among them.

In summary, ontology in philosophy is a method for classifying and interpreting existence, while in computer science and information science, it serves as a tool that enables software to understand domain-specific knowledge.

2.6.2 The Semantic Web

Since Tim Berners-Lee introduced the World Wide Web (WWW) over twenty years ago (Berners-Lee et al., 1994), it has become a pivotal advancement in information technology and global communication. This revolutionary platform allows users worldwide to access and interact with a vast repository of electronic documents and resources, identified uniquely by Uniform Resource Identifiers (URI). The WWW has grown exponentially, becoming the largest information repository accessible to an estimated 500 million users globally as of 2007 (Bui et al., 2007).

Despite its success, the efficiency of the current web in knowledge sharing is debated. This has spurred interest in Semantic Web technologies, designed to enhance web efficiency and now gaining traction across sectors like bioinformatics, medicine, finance, and construction (Antoniou and Harmelen, 2008;

Warren and Alsmeyer, 2005; Abanda et al., 2013b). The growing popularity of the Semantic Web in various fields underscores the need to explore the limitations of current web technologies to highlight the Semantic Web's potential benefits.

The Semantic Web is a framework designed to enhance data intelligibility and usability on the Web, transforming it to be not only user-friendly for humans but also interpretable and actionable by computer programs. This vision, championed by the World Wide Web Consortium (W3C), involves assigning explicit meanings to data, facilitating interoperability among diverse data sources. This approach allows for data to be seamlessly integrated and automatically processed across platforms. Below are some fundamental concepts of the Semantic Web:

Resource Description Framework (RDF)

RDF is a data model used to represent information about resources on the web. It uses triples (subject, predicate, object) to represent data, allowing the attributes of anything and the relationships between those attributes to be explicitly described. This structure of RDF clarifies the meaning of data, facilitating the sharing and processing of data by different applications.

RDF Schema (RDFS)

RDFS is a language that builds upon RDF to provide enhanced capabilities for describing data. It allows for the definition of relationships between classes and properties, as well as the hierarchical structure of classes, thus offering a richer structured semantic framework for data.

Web Ontology Language (OWL)

OWL is a more powerful language for defining and instantiating ontologies on the web. It enables the creation of more complex representations of knowledge, including relationships between classes, characteristics of properties (such as symmetry, transitivity, etc.), and more complex classification logic. OWL supports sophisticated querying and reasoning, enabling applications of the Semantic Web to process data more intelligently.

SPARQL

SPARQL is a query language and protocol for querying RDF data. It allows developers to write complex queries to extract or manipulate data that meets specific conditions, supporting efficient access and integration of Semantic Web data.

URI and IRI

In the Semantic Web, all resources (including entities, concepts, and relationships) are identified and referenced by unique identifiers, namely Uniform Resource Identifiers (URI) or Internationalized Resource Identifiers (IRI). This ensures the global uniqueness and interoperability of data.

Linked Data

Linked Data is an important concept of the Semantic Web, referring to the use of RDF to connect data dispersed across different data sources. It relies on URIs to provide a globally unique identifier for data entities and uses RDF to describe the relationships between data entities, thus building a global, interconnected data network.

2.6.3 Web Ontology Language (OWL)

OWL (Web Ontology Language) is a standard language used for defining and instantiating ontologies on the web. It is designed to represent rich and complex knowledge about entities, classes, properties, and their interrelationships. OWL allows data to be interpreted and utilized not only by humans but also by computer programs, enhancing the efficiency of data sharing and reuse. As part of the Semantic Web technology stack, OWL aims to make web content more accessible to machines, thereby supporting more complex queries, reasoning, and knowledge management.

2.6.3.1 Key Features of OWL

Expressive Power: OWL provides richer expressiveness compared to XML, RDF (Resource Description Framework), and RDFS (RDF Schema). It can represent highly complex relationships and constraints, such as equivalent classes, inverse properties, enumerations of members, transitivity of properties, and restrictions on cardinality and scope.

Logical Reasoning Support: With ontologies defined in OWL, reasoning engines can perform complex logical reasoning, such as determining class membership, checking data consistency, enabling automatic classification, and deriving new knowledge.

Interoperability: As a W3C standard, OWL enhances interoperability between different systems and applications, allowing data and knowledge from diverse sources to be integrated and reused.

2.6.3.2 Versions and Variants of OWL

OWL DL: Offers full logical reasoning support, ensuring decidability (i.e., reasoning processes complete in finite time). It is based on Description Logic, a formal logic system used for knowledge representation.

OWL Lite: Supports simple class hierarchies and simple constraints, designed to meet the needs of users who do not require the full expressiveness of OWL DL.

OWL 2: The latest version of OWL, which includes several new features and capabilities, such as new data types, richer property expressions, pattern matching, and several compatible sublanguages (profiles) to meet different application needs and performance considerations.

OWL provides a powerful language for defining and manipulating complex knowledge structures and is a key technology for building Semantic Web applications and achieving data and knowledge sharing and reuse.

2.6.4 Semantic Web Rule Language (SWRL)

Semantic Web Rule Language (SWRL) is a rule language based on the Semantic Web technology stack, used for expressing rules about ontologies composed of OWL (Web Ontology Language). In this way, SWRL enhances the expressiveness of OWL, allowing users to define complex logical rules to deduce, constrain, or generate new knowledge about entities within ontologies. SWRL aims to integrate and extend the existing Web Ontology Language and rule systems' semantics, offering a unified framework to express logical rules (Eiter et al., 2008).

2.6.4.1 Basic Components of SWRL

SWRL rules consist of an antecedent (if part) and a consequent (then part), both of which are composed of lists of atoms. These atoms can represent class memberships, property values, equality, inequality, etc. The general form of a rule is:

$$antecedent \Rightarrow consequent$$

If the antecedent is true, then the consequent is also considered true. This allows for the deduction of new facts from existing information in ontologies

2.6.4.2 Types of SWRL Atoms

SWRL encompasses several types of atoms, including:

- Class Atom: Indicates that an instance belongs to a specific class.
- Property Atom: Represents the relationships between properties of instances.
- Equality Atom: Asserts that two instances are equal.
- Inequality Atom: Asserts that two instances are not equal.
- Built-in Atom: Utilizes SWRL's built-in function library to express more complex conditions and computations.

2.6.4.3 Applications of SWRL

SWRL is extensively used in areas such as data integration, knowledge reasoning, Semantic Web services, and information retrieval. It allows developers to define more complex business logic and knowledge reasoning rules on the basis of OWL ontologies, thereby facilitating the creation of more intelligent applications. For example, SWRL can be used for automatic classification, data consistency checks, and inferring relationships between entities that are not explicitly defined.

2.6.4.4 Limitations of Using SWRL

While SWRL is powerful, it also has some limitations. The most significant is that the expressiveness of SWRL can lead to undecidability in reasoning processes (i.e., it cannot guarantee completion of reasoning within finite time), depending on the reasoning engine used and the specific structure of the ontologies. Moreover, the management and maintenance of SWRL rules require careful consideration of potential conflicts and complexities among rules.

Overall, SWRL is a potent tool for enhancing the expressiveness and reasoning capabilities of ontologies in the Semantic Web domain. However, its use should be carefully considered based on the specific application context and available performance.

2.6.4.5 Role of SWRL

SWRL is used to define additional logical rules within OWL ontologies, which can be executed by reasoning engines such as Pellet or HermiT to automatically derive knowledge that is not explicitly declared within the ontologies. In this example, SWRL rules are utilized to automatically calculate the cross-sectional area of a column based on its width and height. Such rules make knowledge representation more dynamic and flexible, enhancing the practicality and expressive power of the ontology. SWRL, written in conjunction with OWL (Web Ontology Language), is designed to enhance OWL's expressiveness by defining rules that deduce knowledge and describe relationships between classes and properties within ontologies.

2.7 Summary of Literature Review

The literature review highlights the growing adoption of Building Information Modelling (BIM) in the construction industry. The increasing repository of BIM models within various organizations has resulted in a significant accumulation of data, including information on geometry, materials, energy, and safety parameters. Despite the richness of this data, it has traditionally been underutilized, with many models archived without fully exploring their potential, thus representing an underleveraged asset. The opportunities to harness the embedded data, information, and knowledge within BIM models for organizational benefit are vast; however, there remains a distinct gap in effectively leveraging this data for tangible outcomes. Recent advancements in machine learning have opened new avenues for exploiting BIM data, as demonstrated by Hu & Castro-Lacouture (2018), who noted the successful application of these methods in clash detection. The agility and efficiency of artificial intelligence (AI) methods, which rely on navigating pre-trained solutions, mark a significant departure from conventional approaches. Chen (2019) suggests that machine learning outputs more accurately reflect actual building performance compared to traditional simulation methods. By utilizing training datasets derived from real historical data or simulated mock data, machine learning methods address the time-intensive nature of conventional energy simulations, which are often prone to biases.

BIM’s integration of energy simulation data into its models provides comprehensive lifecycle datasets, potentially enabling automated energy management throughout a building's lifespan. This integration between BIM and machine learning algorithms addresses specific challenges in building design and performance. However, the balance between energy consumption and occupant comfort levels remains largely unexamined. Traditional energy and comfort assessment methods, which depend on continuous data collection from strategically placed sensors, are not only costly but also susceptible to disruptions and inaccuracies. Moreover, while simulation software can depict a building's operational status, it often requires model reconstruction for each analysis, thus hindering real-time performance prediction capabilities.

The review also explores the multidimensional nature of building comfort, including thermal, acoustic, visual, and air quality factors. These dimensions are individually important and collectively influential, affecting occupant perception and building energy efficiency in complex ways. The literature reveals a critical gap—the absence of a comprehensive comfort framework that can be effectively integrated during the design phase. Although extensive research has been conducted on measuring and enhancing individual comfort dimensions, few studies have focused on integrating these dimensions from the onset of building design. This shortfall limits the holistic consideration and optimization of occupant comfort and building performance during the planning and implementation stages.

Further examination of the intersection of BIM, machine learning, and building energy and comfort uncovers an analytical gap. While BIM has the capacity to capture extensive comfort-related data, the absence of a robust analytical framework prevents the translation of this data into actionable design strategies and decisions. Traditional methods for assessing energy and comfort provide only operational snapshots of building performance, often lacking the capacity to anticipate long-term or dynamic changes. To bridge this gap, the study proposes developing a BIM-based machine learning engine prototype, which will be tested and validated through case studies in subsequent chapters. The goal of this engine is to integrate and enhance various dimensions of comfort while predicting and improving overall building performance from the design stage. This innovative approach has the potential to transform building design and management by providing a holistic framework for addressing both energy efficiency and occupant comfort.

To further address the complexities of integrating and managing comfort dimensions, the review incorporates ontology as an essential tool. Ontology offers a structured way to organize knowledge, representing the relationships between different comfort factors and enabling advanced reasoning capabilities. By integrating ontology within the BIM-ML framework, the system can dynamically adapt to changing environmental and occupant conditions, thus facilitating real-time comfort adjustments and decision-making. The use of ontology enhances the interoperability and scalability of the system, supporting more sophisticated queries and enabling a seamless integration of BIM, ML, and comfort assessment components.

In conclusion, the literature review establishes a foundational link between BIM, machine learning, and building comfort and energy management. It highlights the potential for integrating these technologies into an intelligent building design and management system. This integrated approach, supported by ontology, sets the stage for developing a cohesive, intelligent, and adaptable system for building design and management, which will be further explored and validated through case studies in the subsequent chapters. This advancement represents a crucial step toward a more holistic and efficient approach to sustainable building design, where energy efficiency and occupant comfort are simultaneously optimized.

3 Methodology

3.1 Overall Research Design

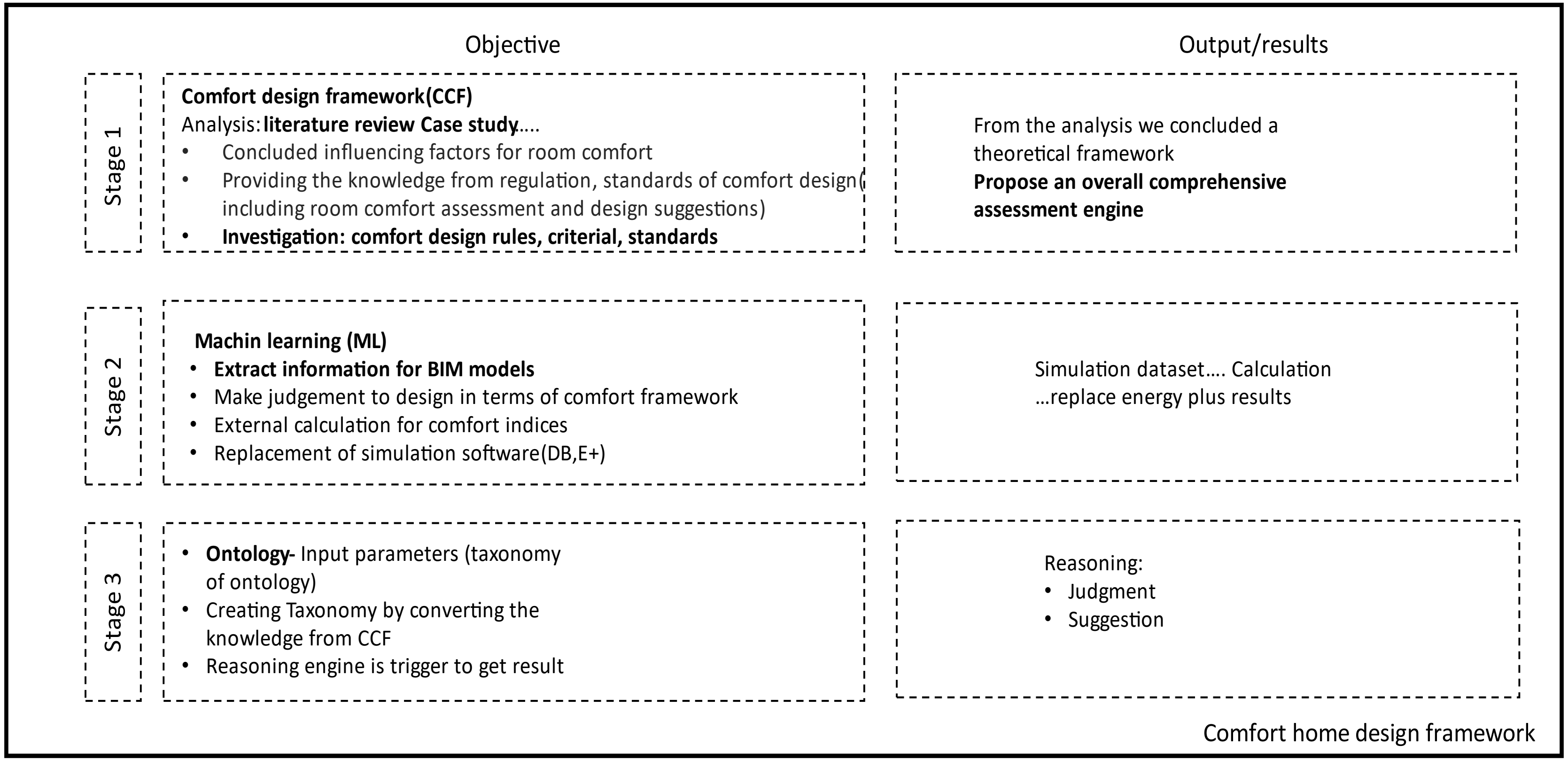
The methodology of this study is structured across three key chapters, each addressing a critical aspect of the comprehensive building comfort design framework and its implementation using machine learning and ontology. The approach integrates theoretical investigation, literature review, practical application of machine learning algorithms, and the development of an ontology-based comfort assessment system. These steps ensure a holistic and efficient evaluation and prediction of indoor environmental comfort.

The overall research design, as illustrated in Figure 3.1 and 3.2, are both built around three main components:

- Comprehensive Comfort Framework (CCF): Developed through a combination of quantitative and qualitative research methods, the CCF provides a unified metric for assessing building comfort. It incorporates four dimensions: thermal comfort, acoustic comfort, visual comfort, and air quality. The CCF simplifies the assessment of comfort by combining these factors into a single metric, facilitating decision-making and collaboration during the design process.
- Machine Learning (ML): ML algorithms replace traditional energy simulation models, allowing for more efficient and accurate predictions of comfort and energy performance. The machine learning process involves data collection, preprocessing, feature engineering, model training, and validation, ensuring that comfort assessments are both dynamic and data-driven.
- Ontology: The ontology framework structures the knowledge gained from the CCF, enabling the reasoning engine to perform real-time comfort assessments. The reasoning engine connects to external machine learning models to make dynamic adjustments to building systems, ensuring that comfort conditions are maintained in real-time.

By integrating these three elements, the methodology provides a comprehensive and adaptive approach to assessing and improving building comfort, offering significant advancements over traditional methods.

Plan stages in whole research



Objective

Output/results

Stage 1

Comfort design framework(CCF)

Analysis: **literature review Case study**….

- Concluded influencing factors for room comfort
- Providing the knowledge from regulation, standards of comfort design(including room comfort assessment and design suggestions)
- **Investigation: comfort design rules, criterial, standards**

From the analysis we concluded a theoretical framework

Propose an overall comprehensive assessment engine

Stage 2

Machin learning (ML)

- **Extract information for BIM models**
- Make judgement to design in terms of comfort framework
- External calculation for comfort indices
- Replacement of simulation software(DB,E+)

Simulation dataset…. Calculation …replace energy plus results

Stage 3

- **Ontology-** Input parameters (taxonomy of ontology)
- Creating Taxonomy by converting the knowledge from CCF
- Reasoning engine is trigger to get result

Reasoning:

- Judgment
- Suggestion

Comfort home design framework

Figure 3. 1 Three components of comprehensive building comfort framework

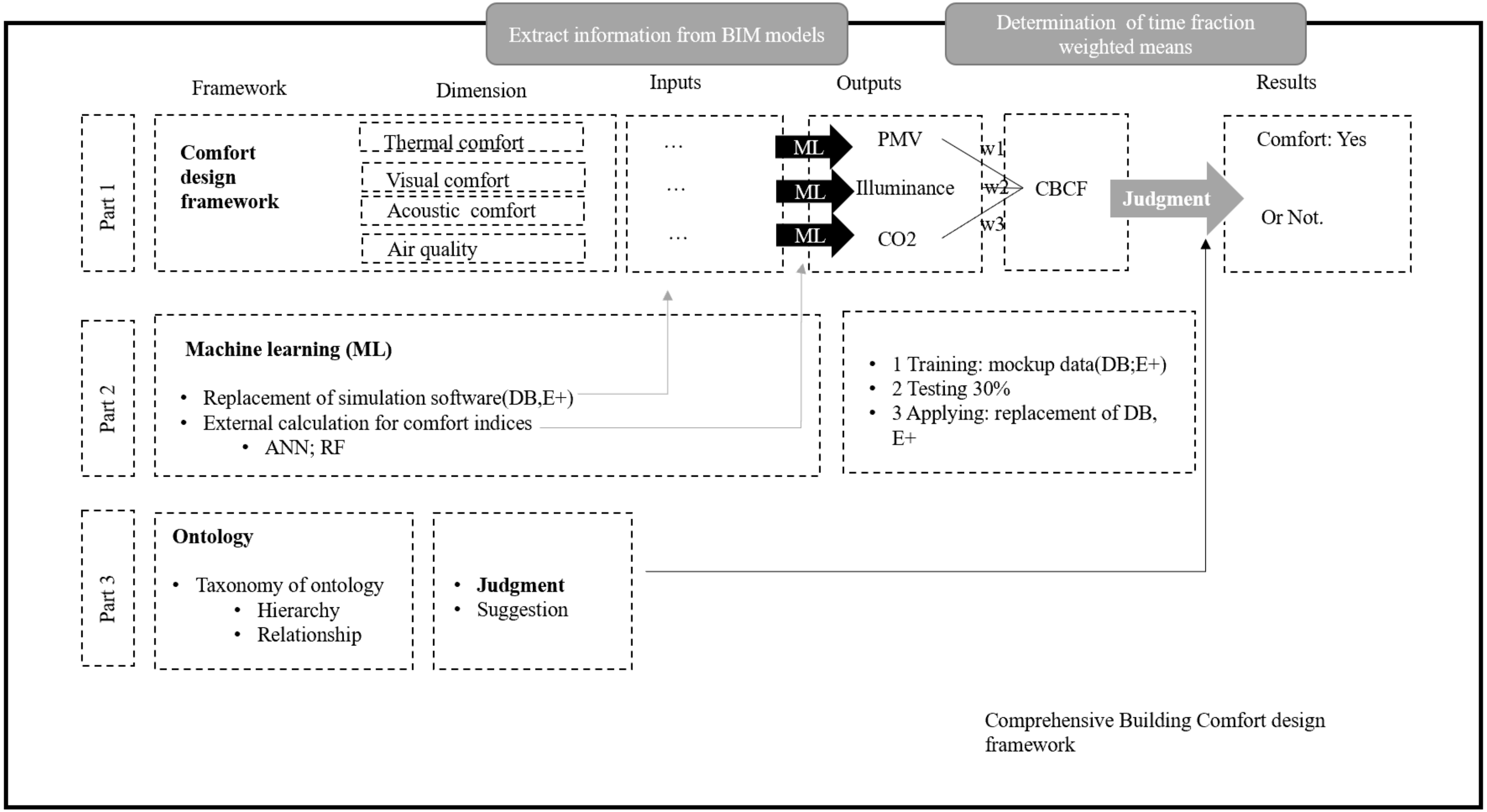


Figure 3. 2 Comprehensive building comfort framework and their interconnections

3.2 Chapter 4: Comprehensive Building Comfort Design Framework

The foundation of the methodology is laid out in Chapter 4, where a Comprehensive Building Comfort Design Framework is developed. This framework was developed through a combined quantitative and qualitative research approach, examining legal and regulatory standards to create a unified comfort metric. It is rooted in a thorough investigation of existing literature, industry standards, and comfort regulations. These resources helped define and formalize the key factors influencing indoor comfort: thermal comfort, acoustic comfort, visual comfort, and air quality. It integrates four key dimensions into a single comfort index for holistic building assessments. The methodology, presented in Chapter 3, defines specific comfort indicators and uses normalization to provide an overall comfort measure, moving beyond traditional thermal comfort assessments. This approach simplifies complex comfort variables, making it easier for stakeholders to evaluate and select optimal building designs, enhancing collaboration and decision-making in the design phase.

The framework integrates these factors into a unified comfort index, allowing for a holistic assessment of a building's indoor environment. The chapter also defines key variables and metrics, such as the Predicted Mean Vote (PMV) and Predicted Percentage of Dissatisfied (PPD), which are used to evaluate comfort levels. Through this comprehensive analysis, the study identifies the limitations of traditional comfort assessment methods and sets the stage for a more dynamic and adaptable approach.

3.3 Chapter 5: Machine Learning for Comfort and Energy Prediction

Building on the comfort framework established in Chapter 4, Chapter 5 introduces machine learning (ML) techniques to predict both building energy consumption and room comfort more effectively. By leveraging large datasets generated from building simulations and real-time sensor data, machine learning algorithms, such as Linear Regression (LR), Artificial Neural Networks (ANN), and Random Forests (RF), are employed to model the complex relationships between comfort factors and environmental conditions.

This chapter highlights the use of ML in predicting daylight illuminance, thermal comfort indices (PMV and PPD), and discomfort hours. These predictions enable faster, more accurate evaluations of comfort conditions compared to traditional simulation-based methods. The chapter also details the training, validation, and performance evaluation of the machine learning models, emphasizing the importance of data preprocessing, feature selection, and algorithm comparison in achieving accurate predictions. By automating the prediction process through machine learning, the study significantly improves the speed and efficiency of comfort assessments, allowing for real-time adjustments to building systems.

The key steps involved in implementing a machine learning process are as follows:

- Data Collection: Gathering relevant data for the problem at hand.

- Data Preprocessing: Cleaning and preparing the data for analysis.
- Feature Engineering: Extracting and selecting important features from the dataset.
- Model Training: Training the model using the prepared dataset.
- Model Evaluation: Validating and fine-tuning the model for optimal performance.
- Result Analysis: Analysing the outcomes and interpreting the results.

These steps form the foundational workflow of a machine learning project, ensuring a structured approach from data acquisition to model deployment.

3.4 Chapter 6: Ontology-Based Comfort Assessment Framework

In Chapter 6, the methodology takes a step further by introducing ontology technology for integrating and managing the diverse factors influencing building comfort. The ontology serves as a formal representation of knowledge within the comfort assessment system, enabling structured data management and logical reasoning. It provides a dynamic interface that connects to external machine learning engines for real-time comfort calculations.

The ontology-based framework encapsulates the key comfort dimensions—thermal, acoustic, visual, and air quality—and their interdependencies. Using Semantic Web Rule Language (SWRL), rules are created to evaluate the overall comfort index (CBC) by combining individual comfort factors and their relative importance ($\alpha1$ for thermal, $\alpha2$ for acoustic, $\alpha3$ for air quality, and $\alpha4$ for visual comfort).

Furthermore, the study employs customized comfort preference profiles by using the Analytic Hierarchy Process (AHP) to weight these factors according to user-defined preferences. By assigning weights to each comfort dimension based on personal or contextual priorities, the framework enables tailored comfort assessments for different building types or user groups.

The ontology management system, implemented using Protégé, allows for the continuous modification and expansion of the comfort model as new data or rules are introduced. The rule engine integrates with the machine learning models to dynamically adjust comfort calculations based on real-time environmental data and user feedback. This framework ensures the system is adaptable, scalable, and capable of evolving as new technologies or comfort standards emerge.

3.5 Integration of Machine Learning and Ontology for Comprehensive Building Comfort Frameworks

The final methodological step integrates the machine learning models developed in Chapter 5 with the ontology framework outlined in Chapter 6. This hybrid approach combines the data-driven predictive capabilities of machine learning with the structured, rule-based reasoning of the ontology. The result is a real-time Comfort Assessment Engine capable of dynamically evaluating indoor comfort, predicting future comfort levels, and providing recommendations for system adjustments.

By continuously updating the system with real-time data and user feedback, the methodology ensures that building systems—such as HVAC, lighting, and ventilation—are optimized for energy efficiency and occupant comfort. This integrated system not only enhances the speed and accuracy of comfort assessments but also makes the process more adaptable to changing environmental conditions and occupant needs.

This Figure 3.3 below illustrates the interaction between three main components in the Comprehensive Comfort Framework (CCF): input parameters (taxonomy of ontology), machine learning (ML), and reasoning. Each element plays a crucial role in the system:

- Comprehensive Comfort Framework (CCF):
    - Concludes the key factors that influence room comfort, based on existing knowledge from regulations, standards, and research.
    - Provides the fundamental knowledge needed for comfort assessments and design suggestions, serving as the overarching model guiding comfort evaluation.
- Machine Learning (ML):
    - Acts as an external calculation engine to compute comfort indices based on the input parameters.
    - Replaces traditional simulation software like DesignBuilder (DB) and EnergyPlus (E+), improving the speed and accuracy of comfort predictions.
- Ontology:
    - Organizes the knowledge from the CCF into a structured taxonomy, enabling the system to store and access data logically.
    - The reasoning engine within the ontology processes these parameters and triggers results based on the ML computations and predefined rules.

This integration of machine learning and ontology-based reasoning allows for dynamic comfort assessment and real-time predictions, enhancing the precision and adaptability of building comfort evaluations.

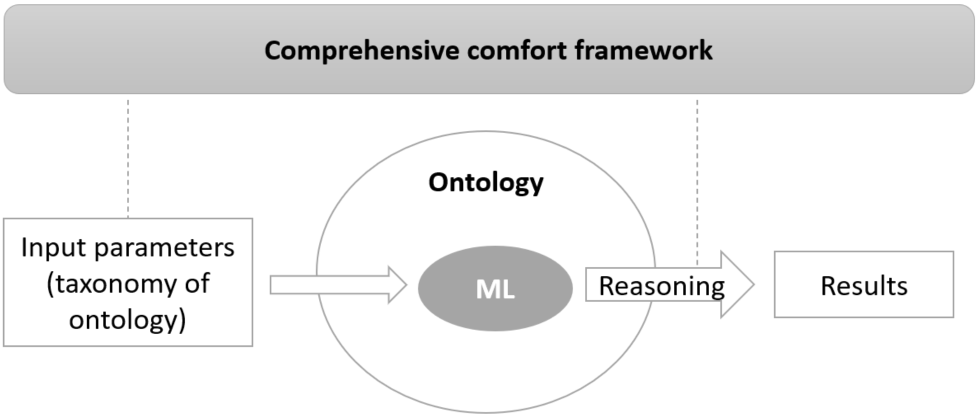

Comprehensive comfort framework
Ontology
Input parameters (taxonomy of ontology)
ML
Reasoning
Results

Figure 3. 3 Interaction between three main components in the Comprehensive Comfort Framework

3.6 Conclusion

The methodology adopted in this study represents a multi-layered approach to improving indoor comfort assessment. Chapter 4 lays the groundwork with a comprehensive comfort framework, Chapter 5 introduces machine learning for rapid comfort predictions, and Chapter 6 integrates ontology to create a flexible, user-driven comfort assessment system. Together, these components form a robust methodology for achieving personalized, real-time comfort optimization in modern buildings, offering a significant advancement over traditional assessment methods. This integrated approach is particularly valuable for smart building applications, where real-time data and adaptive systems are critical for maintaining optimal comfort and energy efficiency. This methodology chapter provides a structured and methodologically sound framework for the research, offering a significant contribution to the fields of BIM, machine learning, and building comfort. It establishes a clear path for future research and development, encouraging continued exploration of these technologies for smarter, more sustainable buildings.

4 Comprehensive and Normalised Building Comfort Design Framework

4.1 Introduction

Humans spend approximately 90% of their time indoors (Ganesh，2021). With the advent of the internet and the ever-changing social environment, more people choose to shop and work from home on their computers or mobile phones. Such a trend is said to increase efficiency and is predestined to create more demand for the quality of the indoor environment. Indoor air conditions can directly affect physical and mental health and productivity, and the choice of systems for building design and air conditioning systems. Based on this reasoning, built environment and human comfort rapidly develop into a combination of multidisciplinary studies involving thermology, architecture, public health, physiology and hygiene study.

Comfort is closely related to well-being, which was defined by Dodge et al. (2012) as "...... when individuals have the psychological, social and physical resources they need to cope with specific psychological, social and physical challenges". In the built environment design industry, the comfort of a room should be an essential thing to consider for both designers and occupants. However, it is important to note that the concept of comfort itself is challenging according to contemporary understandings. Therefore, one of the most important aspects to consider when designing a building is the extent to which it can provide an environment that is comfortable for its occupants. Comfort in the built environment is influenced by a large number of different factors which, if not handled correctly, can lead to discomfort and even harm to mental or physical health (Ganesh，2021). Comfort, of course, includes different aspects, which are classified according to human senses: thermal comfort, indoor air quality, visual comfort, acoustic comfort, etc.

According to conventional belief, the most common studies on architectural housing design tend to revolve around thermal comfort. Although thermal comfort referred to in these studies is only a thermally neutral condition based on theory, it is of general significance that can affect the comfort level of an indoor space. Several other studies have also looked at visual comfort and air quality, among others. The comfort requirements of different people or rooms can vary depending on the individual occupants and the room's functionality. For example, a library or study room may require high visual comfort, whereas a conference room or classroom requires high air quality; a bedroom or living room, for instance, may require high thermal comfort.

This chapter is to develop a framework that can incorporate comprehensive comfort considerations from the user's view to guide architects to improve building comfort design to meet as many requirements as possible for the generic occupant or the situation where the user personalises the level of comfort of the indoor space.

4.2 Building Comfort Design Considerations

The primary function of a building is to provide a comfortable habitat for its occupants by creating a separation between indoor and outdoor environments. Consequently, the design of a building must take into account a range of factors, including structural, architectural, environmental, technical, socio-cultural, functional, and aesthetic considerations. Additionally, the goal of building comfort design is to establish comfortable indoor conditions while ensuring energy efficiency. Achieving this balance requires the careful determination of appropriate values for design parameters that influence indoor comfort.

The parameters that affect building comfort design can be categorized into three groups: building fabric, environmental conditions, and human behaviour. The key parameters within each group are outlined below.

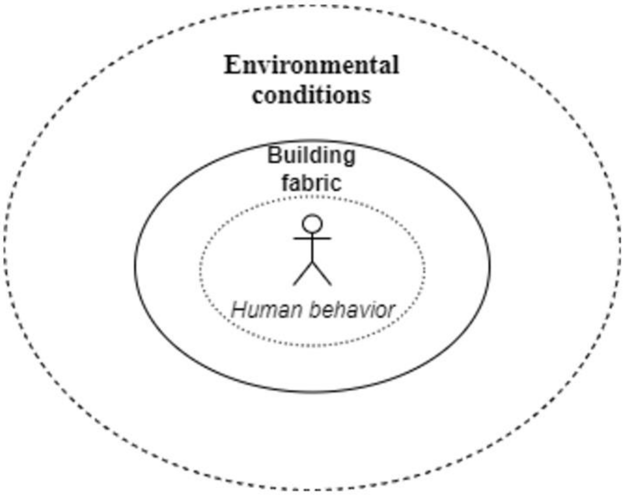


Figure 4. 1 Three divided groups of main parameters in building comfort design

As shown in the Figure 4.1 above, human comfort is protected and influenced by three layers of barriers. The outermost layer is the environmental climate, which objectively impacts human comfort. For instance, the atmosphere protects humans from solar radiation in space, acting as a natural shield. This represents the most external barrier affecting human perception and is an uncontrollable factor. In contrast, the most direct influencing factor is human behaviour. Human behaviour not only includes simple activities like sitting or running but also individual habits, such as what clothes one wears or the thickness of blankets used. These choices can significantly influence perceived body temperature.

Unlike the objective environmental conditions, this layer is subjective and controllable, allowing individuals to adjust their behaviour to modify their comfort level.

Given these two factors, humans have built buildings as a more efficient "outer coat" to enhance comfort. Buildings serve multiple purposes, such as providing warmth in winter, shading in summer, and sheltering from rain, thereby creating a more suitable living environment. Thus, buildings provide humans with additional means to improve comfort under uncontrollable climatic conditions and controllable behaviour adjustments. Therefore, it is crucial to consider user comfort during the building design phase. By integrating climate, behaviour, and building characteristics in the design process, buildings can be tailored to meet human comfort needs more effectively, balancing energy consumption with comfort levels and achieving better overall building performance and user experience.

4.2.1 Building Fabric

The design of building fabrics refers to the structures and components that are intentionally designed and constructed by humans. These elements can be adjusted and modified based on the scale of the building and the prevailing environmental conditions. The parameters attributed to this group can be considered based on a building, a room or an element. The main design parameters related to building comfort and impact on the control of heat, light, sound, air quality, and energy efficiency are partly shown as follows.

a) Design parameters on the building scale
   - Orientation of the building.
   - Position of the building relative to the noise source.
   - Position the building according to the other buildings and the noise source.
   - Building form.

b) Design parameters on the room-scale
   - Position of the room within the building.
   - Dimensions of the room and its shape factor.
   - Orientation of the room.
   - The room's absorption coefficient for solar radiation enters through the transparent component.
   - Sound absorption coefficients of the surfaces inside the room.
   - The total sound absorption coefficient of the room.
   - Light reflectivity coefficients of the surfaces inside the room.

4.2.2 Environmental Conditions

Environmental parameters from the perspective of environmental conditions have a decisive influence on the outdoor environment. Therefore, the parameters of this group are divided into two parts: natural and settlement parameters.

a) The natural parameters are related to the location of the building and the local weather. They are beyond the designer's control, with their given values considered not adjustable. They include:
   - Outdoor air temperature.
   - Solar radiation.
   - Outdoor humidity.
   - Outdoor wind speed.
   - Outdoor illumination levels.
   - Outdoor sound levels.

b) Parameters on the settlement area are usually considered by the designer and can be adjusted or altered, which include:
   - Dimensions and orientation of external obstacles.
   - Solar radiation reflectivity of surrounding surfaces.
   - Light reflectivity of surrounding surfaces.
   - Soil cover and ground nature (plant cover and groups of trees).

4.2.3 Human factor

Human factors, such as clothing and activities, can affect the level of comfort in a building. Such factors may include:

- Age.
- Gender.
- Level of health.
- Clothing worn.
- Type of activity and level of intensity.
- Access to food and drink.
- Acclimatisation.
- Psychological state.
- For example, older people tend to feel cold more often and require additional warmth than younger people.

However, human factors are not within the control of designers and are inherently unpredictable. Nonetheless, they should be considered to some extent during the design phase.

4.3 The Proposed Comprehensive Building Comfort Framework (CBCF)

Generally speaking, human comfort is determined by a combination of factors such as temperature, light, radiation, humidity, airflow and quality and vibrations of the building, as well as noise, odours and other factors specific to life. Among these factors, thermal comfort is the most common factor to be considered in building design. Typically, the feeling of comfort depends on the thermal factors (determined by air temperature, humidity. ventilation, clothing and physical activity) should be referred to as thermal comfort. In recent studies, only thermal comfort has been evaluated with some useful indices such as PMV (Predicted Mean Vote), PPD (Predicted Percentage of Dissatisfaction) being proposed. However, results show that noise and air quality can also affect human comfort. Visual comfort has also been considered to be important as well. Hence, the concept of broad comfort level should be developed to assess the overall state of the indoor environment. A comprehensive comfort framework has been proposed. Under this framework, a new index was proposed to describe the general comfort of humans in an indoor environment and comfort framework.

To ensure the health and productivity of the occupants, the building structure must be adjusted to provide the required thermal, visual, acoustic and air quality conditions for comfort. Therefore, building comfort design should be the basis for designing rooms and assessing their performance. For instance, suppose the performance of a given building design is evaluated, and the comfort conditions are not met. In this case, the values of the relevant parameters should be changed during the design process to ensure that the building comfort design provides thermal, visual and acoustic comfort for the user to meet the aforementioned conditions.

Another objective must be to minimise energy consumption and subsequent expenditure. Under such requirement the building structure should be considered as an integral part of the passive system, with optimum control of heat, light, sound, and air quality. Such a building structure design can increase the performance of its passive systems, thereby reducing the load on the active systems. We aim to develop a comprehensive design framework for building comfort that will assist architects in making holistic decisions in early design stages. There are three main parts in this study, while the comprehensive comfort design framework as part 1 is the basement of the whole research. The other parts will be shown in the next chapters accordingly.

4.3.1 Categories of Indoor Environmental Quality (IEQ)

In general, default input values are used in the absence of national regulations. The default criteria are given for several categories. The design criteria for the indoor environment should be documented together with the prerequisites for the use of the space. Four specific categories are defined in the criteria.

Default input values are given for the indoor environmental quality for each different category. A short description of these categories is given in Table 4.1 (EN 16798, 2019).

Table 4. 1 Categories of indoor environment quality

Category	**Level of expectation**	**Explanation**
IEQ I	High	A high level of expectation, is recommended for spaces occupied by very sensitive and fragile persons with special requirements (handicapped, sick, elderly persons and very young children)
IEQ II	Medium	Normal level of expectation should be used for new buildings and renovations
IEQ III	Moderate	An acceptable, moderate level of expectation, may be used for existing buildings
IEQ IV	Low	Values outside the criteria for the above categories. This category should only be accepted for a limited part of the year

4.3.2 Considering Indoor Comfort Aspects in CBCF

Building can be influenced by environmental factors such as heat, light and sound as well as air quality resulting in different comfort aspects for the user. The primary function of the building in terms of physical environmental factors (heat, light, sound, air quality) is to ensure that:

- Thermal comfort is ensured by controlling the influence of climatic factors.
- Visual comfort by controlling natural light.
- Acoustic comfort by reducing noise as much as possible, or to an acceptable level.
- The level of CO2 for the occupants by adjusting ventilation.

Depending on the comfort aspects, energy consumption simulations can be divided into four types shown in table 4.2.

Table 4. 2 Four types of energy consumption simulation process

Comfort types	Influencing comfort parameters	Simulation models

Thermal comfort	Temperatures, surface temperatures of a wall, operative temperatures, humidity ratio, and so on	building thermal simulation models
Acoustic comfort	The wall features and the outdoor noise level characterising the acoustic zone of the building site	prediction models
Visual comfort	Daylight factors and Light plant designing	prediction models
Air quality	Indoor gaseous pollutants (CO2)	prediction models

4.4 Different comfort aspects considered in CBCF

4.4.1 Thermal Comfort

Thermal comfort is defined in BS EN ISO 7730 as ”…that condition of mind which expresses satisfaction with the thermal environment.”, i.e. the condition when the occupant does not have a distinct feeling of being either too hot or too cold (ISO 7730, 2005).

When the above condition is not met, a potential health hazard may occur. In addition, uncomfortable temperatures can cause an adverse impact on productivity, as well as mental or physical conditions. Therefore, building design must provide or allow means to achieve a comfortable indoor climate.

4.4.1.1 Relative parameters and calculation for Thermal comfort

The environmental variables that influence the conditions of thermal comfort include:

- Air Temperature (Ta),
- Mean Radiant Temperature (Tr)
- Relative air velocity (v),
- Water vapour pressure in ambient air (Pa).

The physiological variables that influence the conditions of thermal comfort include:

- Skin Temperature (Tsk),
- Core or Internal Temperature (Tcr),
- Sweat Rate,
- Skin Wettedness (w),
- Thermal Conductance (K) between the core and skin.

Two human factors influencing thermal comfort conditions include the thermal resistance of the clothing (Icl) and the metabolic rate (H/ADu).

$$ADu = 0.202(weight) \times 0.425(height) \times 0.725$$

Using the above equation, an area of 1.8 m2 represents the surface area of an average person of weight 70 kg and a height of 1.73 m (Fanger 1967). In EnergyPlus, this average person area of 1.8 m2 is used for the body surface area of all thermal comfort models.

The humidification of indoor air is usually not required. Indoor humidity has been shown to have limited effects on thermal sensation and perceived air quality in rooms of sedentary occupancy. However, prolonged high humidity in indoor environment can encourage microbial growth, and very low indoor humidity (<15-20%) can lead to dryness and irritation of the eyes and raspatory systems. Requirements for humidity influence the design of dehumidifying (cooling load) and humidifying systems and will influence energy consumption. The criteria depend partly on the requirements for thermal comfort and indoor air quality and partly on the physical requirements of the building (condensation, mould etc) as well. Additional humidity requirements shall be considered for buildings of special functionalities, such as museums, historical buildings, churches and archives, where humidification or dehumidification of indoor air is usually unnecessary to maintain a specifically required level of humidity.

4.4.1.2 Thermal Comfort Models and Indexes

Fanger (1970) defines PMV as the index that predicts or represents the mean thermal sensation vote on a standard scale for a large group of persons for any given combination of the thermal environmental variables, activity and clothing levels. PMV is based on Fanger's comfort equation (Fanger, 1967).

Fanger's Predicted Mean Vote (PMV) model was developed in the 1970s from laboratory and climate chamber studies. In these studies, participants are dressed in standardised clothing and are asked to carry out standardised activities while being exposed to different thermal environments. In some studies, the researchers choose the thermal conditions, and participants are asked to record how hot or cold they feel using the seven-point ASHRAE thermal sensation scale shown in Figure 4.2 (ASHRAE, 2005). In other studies, participants are given control over the thermal environment themselves, adjusting the temperature until they feel thermally 'neutral' (i.e., neither hot nor cold; equivalent to voting '0' on the ASHRAE thermal sensation scale).

PMV is an index that aims to predict the mean value of votes from a group of participants on a seven-point thermal sensation scale. Thermal equilibrium is obtained when an occupant's internal heat production equates to heat loss. The heat balance of an individual can be influenced by the activity they engage in, clothing insulation, and the parameters of the thermal environment. For example, the thermal sensation is generally perceived as better when the occupants have control over indoor temperature (i.e., natural ventilation through the opening or closing of windows).

Within the PMV index, +3 translates as too hot, while -3 would be perceived as too cold. The detail of the index is shown below.

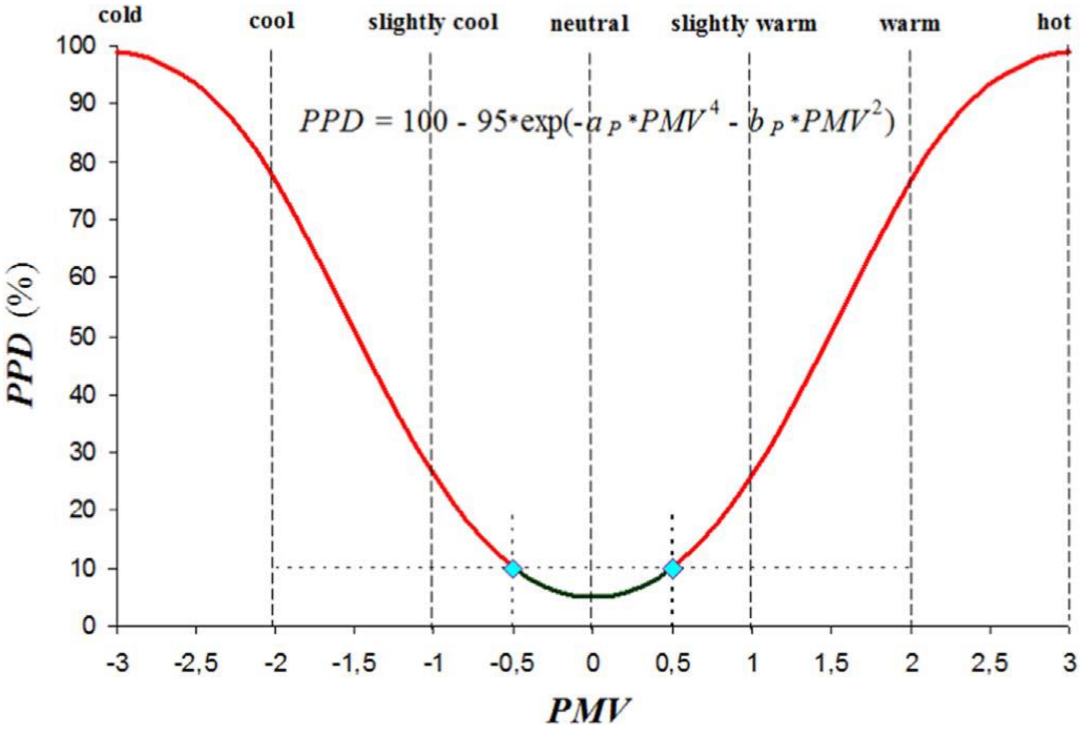


Figure 4. 2 The seven-point ASHRAE thermal sensation scale (ASHRAE, 2005).

Considering the level of satisfaction of the occupants of an indoor space determines the level of thermal comfort, Fanger develops another equation to relate the PMV to the predicted dissatisfaction (PPD) percentage.

Once the PMV is calculated, the PPD, or an index that establishes a quantitative prediction of the percentage of thermally dissatisfied occupants (i.e., too warm or too cold), can be determined. PPD essentially produce a percentage of people predicted to experience local discomfort. The main factors causing local discomfort are unwanted cooling or heating of an occupant's body.

a) The formula for calculating PMV and PPD (ISO 7730, 2005)

- $PMV = 0.303e^{-0.036M} + 0.028\{(M - W)$

$$-0.00305[5733 - 6.99(M - W) - P_s]$$

$$-0.42[M - W - 58.15]$$

$$- 0.000017M(5867 - P_s)$$

$$-0.0014M(34 - \theta_{ai})$$

$$-0.0000000396f_{cl}[(\theta_{cl} + 273)^4 - ([(\theta_c + 273)^4]$$

$$- [f_{cl}h_c(\theta_{cl} - \theta_{ai})]\}$$

- $PPD = 100 - 0.95 * \exp(-0.03353 * PMV^4 - 0.2179 * PMV^2)$
- Comfort Criteria: Occupant comfort is achieved when the PMV value is between -0.5 to +0.5. The corresponding predicted percentage of dissatisfied people falls below 10%.

b) Draft Rating Index (DR)

- $Draft\ Rating\ (\%) = (34-Tx)(Vx-0.05)^{0.62}(0.37\times Vx\times Tu+3.14)$

t=Local air temperature (∘C)

Vx=Local air speed (m/s)

Tu=Turbulence intensity (%)

- Comfort Criteria: Occupant comfort is achieved when the percentage of dissatisfied people due to draft is below 20%.

c) Effective Draft Temperature

- $Effective\ Draft\ Temperature\ (∘F) = (Tx-Tag) -0.07(Vx-30)$

Tx=Local air temperature (∘F)

Tavg=Room average air temperature (∘F)

Vx=Local air velocity (ft/min) or (fpm)

- Comfort Criteria: Occupant comfort is achieved when the effective draft temperature (EDT) is between -3 °F to +2 °F, and the air velocity is less than or equal to 70 fpm.

4.4.1.3 Recommendation of Thermal Comfort Range in Standards

At this stage, target values are first determined for each selected parameter, so that the comfort categories can be objectively separated. Usually, the target values may vary depending on the function of the room. For example, the longer the occupants stay in a living room or a bedroom, the more restrictive conditions are required. Therefore, the representative parameters and the corresponding target values may be different for each indoor environment. However, in order to reduce the complexity of the experiment, it can be assumed that the studied indoor environments are on the same floor and have comparable indoor environments. and that rooms with similar functions are grouped into the same group.

Thermal comfort conditions made by ASHRAE Standard 55 (ASHRAE, 1992) which required four conditions for thermal comfort (CIBSE criteria in the UK):

- Air temperature: for 80% of people, it extends from 20C in winter to 25C in summer.
- Relative humidity: should be above 20% all year, below 60 per cent in summer, and below 80% in winter.
- Air velocity(m/s): in the summer: great asset; in the winter: liability. The comfort ranges from 20 to 60 fpm (0.1 to 0.3 m/s). Noticeable: 60 to 200 fpm (0.3-1m/s). disruptive: above200 fpm (1m/s)
- Mean radiation temperature (MRT): Goal: maintain MRT to ambient art temp. When MRT differs greatly from air temp., its effects must be considered

The recommended PPD-PMV ranges are given in Table 4.3, based on the four indoor environmental categories mentioned earlier (EN ISO7730).

Table 4. 3 The recommended PPD-PMV ranges based on the environmental categories

Category	**Thermal state of the body as a whole**	
	Predicted Percentage of Dissatisfied PPD %	Predicted Mean Vote PMV
I	< 6	−0,2 < PMV < + 0,2
II	< 10	−0,5 < PMV < + 0,5
III	< 15	−0,7 < PMV < + 0,7
IV	< 25	−1,0 < PMV < + 1,0

4.4.2 Visual Comfort

Building energy consumption is a critical factor to consider during the design phase. It has been reported that building energy use accounts for 40% of total energy consumption in Europe (Bull et al., 2012). Therefore, predicting building energy consumption is essential for achieving energy conservation and sustainability. A substantial portion of indoor energy consumption is attributed to the use of artificial lighting. To reduce electricity usage, artificial lighting systems are often regulated based on daylight illumination predictions. Given this context, the daylight factor should be a primary consideration in the building design stage to maximize natural light and minimize electricity consumption.

The sun is the sole natural light source capable of providing suitable living conditions for organisms on Earth. Research indicates that exposure to daylight is an effective psychological healing method (Kittler, 2011). In addition to psychological benefits, adequate natural lighting enhances human comfort and performance. Humans have an instinctive preference for natural sunlight over artificial light during indoor activities, particularly during work hours. Exposure to daylight can significantly influence the physical and mental health of occupants by reducing headaches, eye strain, and stress (ibid). Daylight penetration, influenced by the building facade and surrounding environment, can be controlled through careful design of the facade features.

Visual comfort, therefore, is a crucial aspect of building comfort, encompassing the provision of natural light, external views, and the reduction of glare. Effectively monitoring and managing building energy performance is vital to achieving a balance between energy efficiency and occupant comfort.

4.4.2.1 Daylight System in a Building

The design of daylighting is a subjective argument based on urban regulations, building typology, planning and architectural limits, openness proportions, economic desires or occupants and reactions to lighting conditions.

Since daylight is not accessible during the night and its magnitude decreases along the depth of the room, there always be a combination appropriate with artificial lighting and geometric planning.

From a global point of view, the source of daylight in high-latitude regions is noticeable in summer and winter conditions, while at lower latitudes, daylight variations are reduced (Serra, 1998). Therefore, at high latitudes where the daylight levels are quite low in winter, designers aim to redirect daylight into the building from the brightest part of the sky and to the penetration of sunlight. On the contrary, in regions where daylight levels are significantly high during the year, design strategy often emphasizes limiting the amount of incoming light to avoid overheating and glare. The daylight penetration is a collaboration between the building facade and the peripherals.

Adaptability is considered the key feature of daylighting systems to enhance their effectiveness in indoor environments. There has been a lack of consensus on the majority of acceptable indoor illuminance thresholds from most indices and a lack of reliable glare indices in the presence of the sun within the occupant's line of sight. Similarly, many green building certifications propose specific criteria for assessing fields of view.

4.4.2.2 The Related Variables and Indexes for Evaluating Visual Comfort

Designing daylighting systems is not necessarily limited to energy efficiency or power consumption; visual comfort performance is perceived as equally important, as is visual comfort in the form of glare protection, outdoor views or indoor illumination can be important for the occupants (Tabadkani A. et al., 2021). However, visual comfort is a subjective perception in the visual environment influenced by several co-existing variables. It can affect the well-being of the occupants, as defined by European standards, and similar to the thermal comfort principle, these interrelated variables can be divided into psychological, physical and rational aspects. The first two aspects are less easy to measure and are therefore not discussed in detail in this study. There are three relative indexes will be discussed as fellow:

- Illuminance

- Daylight factor
- Other improved metrics

Illuminance

In existing studies, illuminance is the daylight penetration in space, which is a physical measured in lux at a given point of a surface.

$$E_P = \frac{d_\varphi}{d_A}$$

E_P: lux at giving point P of a surface

d_φ: Incident luminous flux

d_A: The surface area on a diminutive surface around point P

As derived from the above equation, a primary limitation is the difficulty in distinguishing the source of light, whether natural or artificial. Traditionally, data collection on daylight illuminance involves placing sensors at reference points within buildings to gather continuous data. However, these sensors represent an additional cost and are often subject to displacement from the reference point, compromising the accuracy of daylight illuminance measurements.

Furthermore, while simulation models can predict the operational status of a building, they require the creation of a new model for each simulation. This limitation prevents the implementation of real-time illuminance prediction.

Daylight factor (DF)

The illuminance indicator shows the level of daylight at a point in each hour but does not necessarily explain the level of daylight in the space for a year. The daylight factor (DF) is therefore widely used to assess the adequacy of daylight coverage. It is defined as "the ratio of the internal illuminance (Ep_{obs}) to the unobstructed external horizontal illuminance (Ep_{unobs}) at a point in a building under a cloudy CIE" as a worst-case scenario (Moon, 1942).

Indices for evaluating daylight quantity：

$$DF = \frac{Ep_{obs}}{Ep_{unobs}}$$

DF: daylight factor

Ep_{obs}: the internal illuminance

EP_{unobs}: unobstructed external horizontal illuminance

Despite the fact that it does not consider the dynamics of the climate as well as the position of the sun. Furthermore, it also ignores climate dynamics, solar position, building orientation and material reflectance reducing calculation time. As a result, some studies have questioned the accuracy of DF in insolation studies (Zomorodian and Tahsildoost, 2019) because of its static behaviour (Tabadkani, A. et al., 2020). Hence the improved dynamic metrics have been working out.

Improved dynamic metric: UDI

UDI is the shortage of useful daylight illuminance, which is defined as the time fraction of analysis points over a year when the indoor horizontal illuminance falls into a specific range (11). Findings reveal that UDI metric is more accurate to quantify the daylight penetration in a given space due to their range and area divisions that make them feasible to apply especially in shared spaces (Tabadkani A. et al., 2021). The value of UDI is a two-tailed metric which has lower and upper thresholds and an acceptable range. It might lead to glare and thermal stress when the value of UDI is higher than the upper threshold.

$$UDI=\frac{\sum i\,(wfi\cdot ti)}{\sum i\,ti}\in[0.1]$$

$$\begin{cases} UDIoverlit \quad with\ wfi=\begin{cases}1 & if\ Edaylight>Eupper\ limit\\ 0 & if\ Edaylight\leq Eupper\ limit\end{cases}\\ UDIuseful\ with\ wfi=\begin{cases}1 & if\ Elower\ limit\leq Edaylight\leq Eupper\ limit\\ 0 & if\ Edaylight<Elower\ limit\vee Edaylight>Eupper\ limit\end{cases}\\ UDIunderlit\ with\ wfi=\begin{cases}1 & if\ Edaylight<Elower\ limit\\ 0 & if\ Edaylight\geq Eupper\ limit\end{cases}\end{cases}$$

4.4.3 Indoor Air Quality

Human comfort can also be affected by the quality of ventilation in an indoor area (Fanger, 1992). Ventilation is necessary for buildings to remove 'stale' air and replace it with 'fresh' air, as well as to prevent overheating by creating air circulation. Poorly ventilated area creates risks for causing sickness and the spreading of diseases. Indoor air quality and ventilation of buildings are evaluated with a representative sample taken from different air handling units and the zoning on the building is based on.

Indoor air quality and ventilation of buildings are evaluated with a representative sample taken from different air handling units, and the zoning on the building is based. The standard approach to indoor air quality is to specify a recommended level of ventilation (outside air), depending on the number of occupants in the space and a contribution depending on the floor area of the space. There is, however, a need for some clarification and new concepts regarding these issues:

- Ventilation for Non-adapted or adapted occupants
- Use of increased CO2 level as indicator
- Air cleaning as a substitute for outside air
- Ventilation effectiveness
- Personalised ventilation

4.4.3.1 The Methods of Controlling Indoor Air Quality

Indoor air quality shall be controlled by the following means: source control, ventilation, and possible filtration and air cleaning. It is assumed that pollutant emissions are constant in each period considered and lead to a constant ventilation air flow rate for each period.

- Source control

The control of emission of non-human pollutants shall be the primary strategy for maintaining acceptable air quality. It is recommended to identify the main sources of pollutants and to eliminate or decrease them by ventilation. In addition, the choice of building materials, surface preparation, maintenance and furniture impacts the non-human pollutant emissions in rooms, spaces and buildings.

- Ventilation

The design ventilation air flow rates shall be used for designing any type of ventilation system, including mechanical, natural and hybrid ventilation systems.

The design requirements for the ventilation air flow rates shall take into account the pollutant emissions rates left after source control with material selection, local exhaust and other means.

4.4.3.2 The Requirement for Different Categories of Indoor Air Quality

- Non-residential buildings

For designing ventilation systems and determining heating and cooling loads the required ventilation rate should be specified in the design documents based on national requirements or using one of the recommended methods (ref).

Designing for different categories of indoor air quality requires different levels of ventilation rate. The categories of air quality can be expressed in various ways (combination of ventilation for people and building components, ventilation per m2 floor area, ventilation per person or according to required CO2 level). The design documents shall document, which method has been used. The ventilation rates for air quality are independent of season. They depend on occupancy, activities indoors (i.e., smoking or cooking), processes (such as copiers in offices, and chemicals in school buildings) and emissions from building materials and furniture.

In the design and operation, the main sources of pollutants should be identified and eliminated or decreased by any feasible means. The remaining pollution is then dealt with by local exhausts and ventilation.

- Residential buildings

Indoor air quality in residential buildings is affected by a series of parameters and sources such as the number of occupants and the duration of their presence, emissions from activities such as cooking or exercising, and emissions from furnishing, flooring materials and cleaning products. Humidity is of particular concern in residential ventilation as most adverse health effects and building conditions such as condensation and moulds are related to humidity. Several of these sources cannot be influenced or controlled by the designer. Required design ventilation rates shall be specified as an air change per hour for each room and/or fresh air supply or required exhaust rates (bathroom, toilets, and kitchens) or given as an overall required air change rate. Most national regulations and codes give precise indications on detailed airflows per room and shall be followed. The required rates shall be used for designing mechanical, natural and exhaust ventilation systems.

4.4.3.3 Calculation The Design Parameters for Indoor Air Quality

Design parameters for indoor air quality shall be derived using one or more of the following methods (EN 16798, 2019):

- Method 1: Method based on perceived air quality.
- Method 2: Method using limit values for substance concentration.
- Method 3: Method based on predefined ventilation air flow rates

There are many relative calculations and indoor air quality indexes based on perceived air quality as follow.

The air cleaning efficiency is calculated as:

$$\varepsilon_{clean} = (CU - CD) \,/\, CU \cdot 100$$

where,

εclean = air cleaning efficiency
CU = gas concentration before air cleaner
CD = gas concentration after air cleaner

The ventilation rates specified in the standards are the required rates at breathing level in the occupied space. Ventilation effectiveness as follows:

$$\varepsilon_v = \frac{Ce - Cs}{Ci - Cs}$$

Where,
Ce = Pollutant concentration in extracted air
Cs = Pollutant concentration in the supply air
Ci = Pollutant concentration at breathing level

Total ventilation rate:

$$V = Vbz/\varepsilon_v$$

Where,

Vbz = breathing zone ventilation
ε_v = Ventilation effectiveness

Indoor air quality percentage dissatisfaction (IAQ):

Index IAQ can be determined by the use of the decipol unit C proposed by Fanger (1988), as:

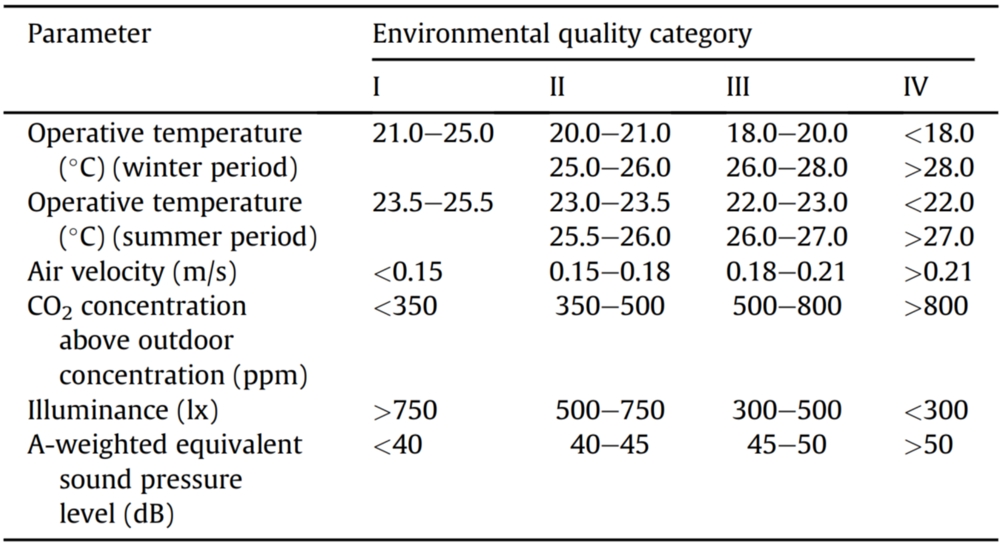
$$IAQ = \exp\left(5.98 - \sqrt[4]{\frac{112}{C}}\right)$$

In the ASHRAE Standard 62 (1984), it is assumed that QPD ≤ 20% is comfortable air quality.

4.4.3.4 The target value of the index of indoor air quality

The ventilation effectiveness depends on the air distribution efficiency and the type and position of the pollution source; therefore, this value is not a system characteristic. In the criteria, the design parameters of air quality are defined in Table 4.4 below.

Table 4. 4 The defined design parameters of air quality



Parameter	Environmental quality category			
	I	II	III	IV
Operative temperature (°C) (winter period)	21.0–25.0	20.0–21.0 25.0–26.0	18.0–20.0 26.0–28.0	<18.0 >28.0
Operative temperature (°C) (summer period)	23.5–25.5	23.0–23.5 25.5–26.0	22.0–23.0 26.0–27.0	<22.0 >27.0
Air velocity (m/s)	<0.15	0.15–0.18	0.18–0.21	>0.21
CO_2 concentration above outdoor concentration (ppm)	<350	350–500	500–800	>800
Illuminance (lx)	>750	500–750	300–500	<300
A-weighted equivalent sound pressure level (dB)	<40	40–45	45–50	>50

The design ventilation rates are calculated based on a mass balance formula for the substance concentration in the space taking into account external concentration. If CO_2 is used as a tracer of human occupancy, the default limit values are extracted from below Table 4.5. The listed CO_2 values can also be used for demand-controlled ventilation.

Table 4. 5 Extracted default limit values if CO_2 is used as a tracer of human occupancy

Category	Expected percentage dissatisfied	Airflow per non-adapted person 1/ (s per person)	Corresponding CO_2concentration above outdoors in PPM for ono-adapted persons
I	15	10	550 (10)
II	20	7	800 (7)
III	30	4	1350 (4)
IV	40	2.5	1350 (4)

4.4.4 Acoustic Comfort

4.4.4.1 Noise Affected Indoor Comfort

The level and source of noise in or around a building can negatively affect comfort of the occupants. Noise nuisance is defined as excessive noise or disturbance that may have a negative effect on health or quality of life. One example of such can be poor soundproofing leading to the occupants being able to hear their neighbours.

For ventilation design, the required sound levels should be specified in the design documents based on national requirements. The noise from the HVAC systems of the building may disturb the occupants and prevent the intended use of space. Requirements should be given for noise from service equipment inside or outside the building assuming windows are closed. Such requirements can be lowered in buildings where the occupants can control windows or noise-generating appliances.

The criteria apply to the sources from the building as well as the noise level from external service equipment. The criteria should be used to limit noise level from mechanical equipment and regulate soundproofing requirements for the noise from outdoors and adjacent rooms.

Ventilation should not rely on opening of windows in areas where high intensity outdoor noise where it is not possible to reach the target level when airing or if the building is located in an area with high outdoor noise level compared to the level the designer wishes to achieve in the indoor zone. National regulations often set requirements for ventilation conditions (including airing).

4.4.4.2 Calculations

ACOU presents the impact of noise on human comfort perception in a building environment. Noise could originate from road traffic, industrial premises, construction or community activities. The discomfort caused by noise can be calculated based on the results of Clausen's study(Clausen et al., 1993). The discomfort of the acoustic equation is as follows:

$$ACOU = 4.35 \int_{-\infty}^{noise\ level} \exp\left(-\left(\frac{x-58.6}{13}\right)^2\right)$$

Where x is the class of noise in dB, in engineering, it is usually assumed that ACOU≤ 20% as a comfortable noise surrounding.

4.5 The CBCF Normalization & Implementation

Because the majority of a building’s environmental impact comes from energy consumed to provide comfort for indoor users, this impact is felt throughout the life cycle of a house and continues to have a cumulative effect on the environment. The most significant of such is the emission of greenhouse gases, which directly contribute towards climate change. It is therefore essential to improve energy performance and reduce the environmental impact of buildings by establishing a framework for indoor comfort and building design optimisation process. The implications of this study are significant and have a considerable potential to mitigate carbon dioxide emissions.

Generally, buildings are planned and constructed to provide a suitable environment for their occupants. Under many functionalities, such as offices where uncomfortable conditions lead to loss of productivity and increased worker stress and discontent, the comfort of the occupants should be of greater importance than energy use. As a result, HVAC equipment is commonly installed in commercial space, even if this result is often unsatisfactory. This runs counter to the global strategy to improve energy efficiency. Therefore, we specify a comprehensive comfort framework to assess the combined energy performance of building comfort systems.

The comfort framework constructed in this research evaluates the overall comfort of indoor environments and the comfort level of the whole building by analysing the different aspects of indoor comfort and the requirements of common standards. This evaluation is achieved through two composite indices: the indoor comfort index and the overall building comfort index. Furthermore, by surveying the individual comfort preferences of specific potential occupants, a comfort signature is added to the knowledge base to further optimise building design.

The comprehensive comfort framework has four realisation steps:

- Step 1: The comfort factors and their parameters are defined.
- Step 2: Indoor comfort is assessed by means of criteria.
- Step 3: Construct a comfort level for the whole building and assess it via means of a combination of indoor comfort from the previous step.
- Step 4: Assess individual comfort signature.

This framework can be used in the initial design phase of a building for comfort assessment and for designers to refer to in order to optimise. However, environmental parameters can only be obtained through digital simulation. Therefore, unlike the operational phase of construction, the environmental parameters can be monitored by sensors and other instruments.

4.5.1 Determination of Parameters in Indoor Comfort Design

Every comfort factor can be analysed by means of representative parameters, consequently, after the comfort factors have been chosen, the procedure of analysis goes on with the selection of measurable or calculable comfort parameters, each of which corresponds to one comfort factor. Figure 4.3 shows the workflow for determining the comfort design parameters. From comfort classification to comfort representation factors to the selection of the corresponding influencing parameters.

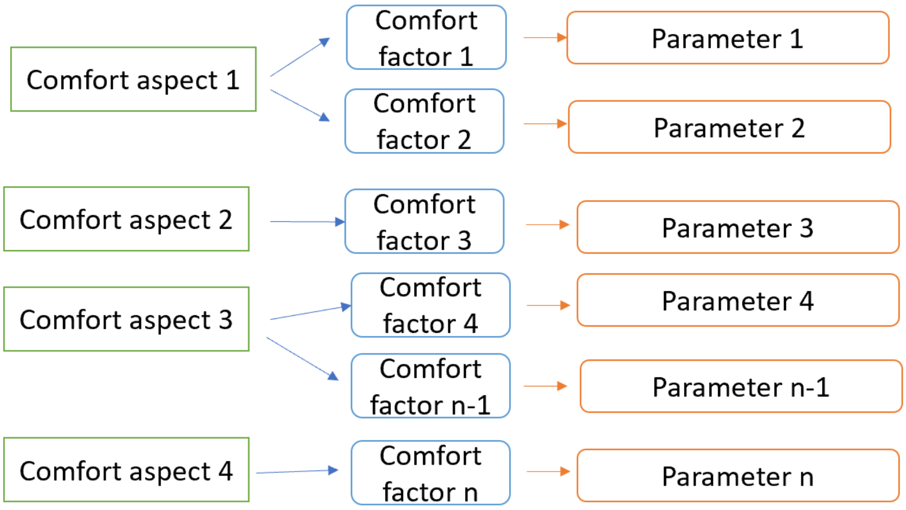


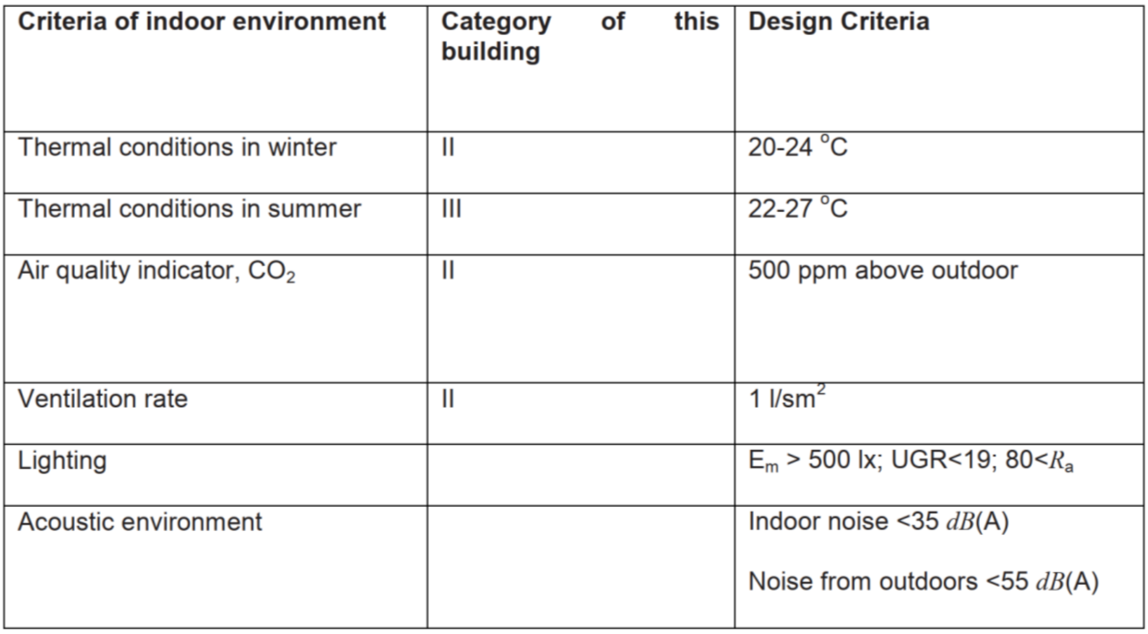
Figure 4. 3 Workflow of determination of parameters in indoor comfort design

4.5.2 Tagert Value of Comfort Indexes in Criteria

At this stage, target values are determined for each selected parameter so that the comfort categories can be objectively distinguished. Typically, the target values may vary depending on the function of the room. For example, the longer occupants stay in a living room or bedroom, the more restrictive conditions are required. Therefore, the representative parameters and their corresponding target values may differ for each indoor environment. However, to simplify the experiment, it can be assumed that the studied indoor environments are on the same floor and have comparable conditions, and that rooms with similar functions are grouped together. Table 4.6 below shows the classification based on criteria for energy calculations.

EN16798 defines the environmental parameters and criteria necessary to achieve the specified energy performance targets. According to EN16798, default design values are provided for each category of performance indexes. Additionally, noise levels generated by continuous building systems are addressed according to the standards outlined in EN12464.

Table 4. 6 Comfort divided categories of indoor environments according to EN15251



Criteria of indoor environment	Category of this building	Design Criteria
Thermal conditions in winter	II	20-24 °C
Thermal conditions in summer	III	22-27 °C
Air quality indicator, CO_2	II	500 ppm above outdoor
Ventilation rate	II	1 l/sm^2
Lighting		E_m > 500 lx; UGR<19; 80<R_a
Acoustic environment		Indoor noise <35 dB(A) Noise from outdoors <55 dB(A)

Table 4. 7 Default design values for each category of the indexes according to EN 16798

		Thermal comfort	**Visual comfort**	**Indoor air quality**	**Acoustic comfort**
Category	Level of expectation	PPD	Illuminance (lux)	CO_2 (ppm)	Acou (%)
I	comfortable	≤ 6	300	≤ 500	≤ 20
II	Slightly uncomfortable	≤ 10		≤ 800	
III	uncomfortable	≤ 15		≤ 1350	
IV	Very uncomfortable	≤ 25		≤ 1350	

4.5.3 Normalisations of Comprehensive Comfort Value

Once the target values for parameters related to comfort limits have been established, the selected comfort parameters and their variations over time are assessed using energy simulation software (e.g., DesignBuilder, EnergyPlus). These simulations need to be conducted over an extended period, such as three months or a full year, to generate a sufficient amount of data. Consequently, the values and fluctuations of each parameter over time are simulated, and this data is further analysed to determine the time scale. The time scale represents the proportion of time during which the calculated values remain within the comfort level range.

The outcomes of this phase are presented in a test score matrix, which provides detailed information for each room, illustrating whether the comfort parameters meet the predefined criteria over the simulated time period.

Time fraction matrix [F]:

$$[F] = \begin{bmatrix} f_{1,\mathrm{I}} & f_{1,\mathrm{II}} & f_{1,\mathrm{III}} & f_{1,\mathrm{IV}} \\ f_{2,\mathrm{I}} & f_{2,\mathrm{II}} & f_{2,\mathrm{III}} & f_{2,\mathrm{IV}} \\ f_{3,\mathrm{I}} & f_{3,\mathrm{II}} & f_{3,\mathrm{III}} & f_{3,\mathrm{IV}} \\ f_{4,\mathrm{I}} & f_{4,\mathrm{II}} & f_{4,\mathrm{III}} & f_{4,\mathrm{IV}} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n,\mathrm{I}} & f_{n,\mathrm{II}} & f_{n,\mathrm{III}} & f_{n,\mathrm{IV}} \end{bmatrix} \quad \text{with } \textstyle\sum_{j=\mathrm{I}}^{\mathrm{IV}} f_{i,j} = 1$$

Up to this point, no hierarchy of comfort aspects has been established. However, in indoor environments, the importance of various comfort aspects varies depending on the tasks performed by occupants and the intended use of the space. For instance, in reading rooms, visual or acoustic comfort may take precedence over thermal comfort, and so on. The generic matrix element ($f_{i,j}$) represents the fraction of time during which the values of the i-th parameter fall within the range limits defining the j-th quality level.

To ensure the general validity of the evaluation procedure, a weight factor (W_i)is assigned to each parameter, reflecting its relative importance. The relative weight (W_i)is then determined using the following relationship:

$$w_i = \frac{w_i}{\Sigma_{j=1}^{N} w_j}$$

These relative weights will be put together to form the Relative Weight Vector, structured with as many rows as the number of selected parameters and that, therefore, can be written as:

Relative Weight Vector {w}

$$\{w\} = \begin{Bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ \cdots \\ w_n \end{Bmatrix}$$

Weights that express a conventional hierarchy of comfort aspects based on the intended use of environments may be sufficient for this task. The weighting of selected comfort parameters, however, requires an examination of the relationship between perceived comfort, environmental characteristics, and the occupants' status or activities. It is important to emphasize that the objective of the proposed procedure is to classify the indoor comfort level to facilitate comparisons of the environmental performance of different buildings. Thus, weights reflecting a conventional hierarchy of comfort aspects, aligned with the intended use of the space, may be adequate.

Time Fraction Weighted Mean Vector $\{f\}$

$$\{\bar{f}\} = [F]^T\{w\}$$

with:

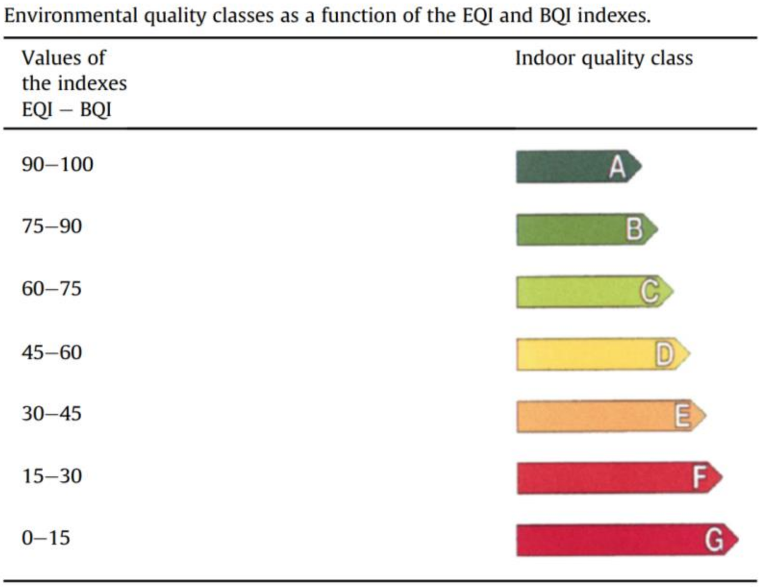
$$\{\bar{f}\} = \begin{Bmatrix} \bar{f}_I \\ \bar{f}_{II} \\ \bar{f}_{III} \\ \bar{f}_{IV} \end{Bmatrix}$$

It can be processed to obtain a single and simple index, namely the Environmental Quality Index, EQI, defined by the following relationship:

$$\mathrm{EQI} = 100\bar{f}_{\mathrm{I}} + 70\bar{f}_{\mathrm{II}} + 35\bar{f}_{\mathrm{III}}$$

4.5.4 Comfort Level

According to Figure 4.4 below, the comfort level can be defined by the value of EQI accordingly.



Environmental quality classes as a function of the EQI and BQI indexes.

Values of the indexes EQI – BQI	Indoor quality class
90–100	A
75–90	B
60–75	C
45–60	D
30–45	E
15–30	F
0–15	G

Figure 4. 4 The comfort level can be defined by the value of EQI accordingly

To reach this goal a seven-point scale, from A to G, has been drawn up, utilising the basic criteria outlined by EN 15217 Standard in energy classification of buildings. With accordance to this standard, in fact, the definition of performance classes is based on two fundamental rules:

The border between classes B and C is the limit value that should be expected from new buildings; 2. The border between classes D and E is the value that should be expected to be reached by approximately 50% of the building stock.

4.6 Calculation of Comprehensive Building Comfort (CBC)

Given that comprehensive comfort perception is influenced by four main factors—thermal, visual, air quality, and acoustic—a new comprehensive comfort evaluation index, termed CBC, is proposed. The CBC index is defined as follows:

$$CBC = \alpha_1 T + \alpha_2 A + \alpha_3 Q + \alpha_4 V$$

Where T, A, Q, and V represent the thermal comfort index, acoustic comfort index, indoor air quality index, and visual comfort index, respectively. The coefficients α_1, α_2, α_3, α_4 are the corresponding weight factors for each comfort aspect. However, there is no established hierarchy of relative importance among these comfort aspects. Therefore, a customised comfort preference system is proposed, allowing the coefficients (α_1, α_2, α_3, α_4) to be adjusted based on specific circumstances. The detailed calculation and determination of these coefficients will be discussed later in accordance with specific scenarios.

Table 4. 8 Comfort Indexes, Calculations, and Comfort Ranges

Comfort Types	Comfort Indexes and Calculations	Comfort Range
Thermal Comfort	T= $PPD = 100 - 0.95 * \exp{(-0.03353 * PMV^4 - 0.2179 * PMV^2)}$	$T \leq 10\%$
Acoustic Comfort	$A = 4.35 \int_{-\infty}^{noise\ level} \exp{(-(\frac{x - 58.6}{13})^2)}$	$A \leq 20\%$
Visual Comfort	$V = DF = \frac{E_{P\,\mathrm{obs}}}{Ep_{unobs}}$	
Air Quality	$Q = \exp\left(5.98 - \sqrt[4]{\frac{112}{C}}\right)$	$Q \leq 20\%$

4.6.1 Customised Comfort Preference

In reality, there is no inherent hierarchy among the four comfort categories, as occupants may prioritize different aspects of comfort based on the room's function. Therefore, during the design phase, the relative importance of each comfort aspect should be evaluated in relation to the intended use of the space and its maximum potential occupancy. This variability highlights the necessity for flexibility in the design process to accommodate diverse user preferences and individual characteristics, as well as the specific functional requirements of the space.

A client may provide a customized ranking of the four comfort aspects based on their specific preferences. For instance, the comfort priorities for one client might be ranked as follows:

Bedroom: Acoustic Comfort > Thermal Comfort > Air Quality > Visual Comfort

Living Room: Visual Comfort > Air Quality > Thermal Comfort > Acoustic Comfort

Alternatively, for another client's requirements, the rankings could be:

Hotel Room: Acoustic Comfort > Air Quality > Visual Comfort > Thermal Comfort

Study Place: Visual Comfort > Air Quality > Thermal Comfort > Acoustic Comfort

The weight assigned to each aspect can then be calculated using the Analytic Hierarchy Process (AHP) method, which allows for the determination of weights according to each specific ranking.

Table 4. 9 List of values of Analytic Hierarchy Process

u_{ij} value	**Meaning**
1	in comparison with u_j, u_i is as important as u_j
3	in comparison with u_j, u_i is a bit important
5	in comparison with u_j, u_i is obviously important
7	in comparison with u_j, u_i is quite important
9	in comparison with u_j, u_i is extremely important
2,4,6,8	medium value of 1-3, 3-5, 5-7, and 7-9
Reciprocal	$u_{ij} = 1/u_{ji}$

Here is step-by-step process for calculating the coefficients α_1, α_2, α_3, α_4:

Consider a four-factor collective u = (u_1, u_2, u_3, u_4) where these factors represent the thermal factor, acoustic factor, visual factor, and air quality factor, respectively. An evaluation matrix of the influence factors can be expressed as:

$$P = \begin{bmatrix} u_{11}\, u_{12}\, u_{13}\, u_{14} \\ u_{21}\, u_{22}\, u_{23}\, u_{24} \\ u_{31}\, u_{32}\, u_{33}\, u_{34} \\ u_{41}\, u_{42}\, u_{43}\, u_{44} \end{bmatrix}$$

To calculate the maximum character roots of the matrix and its character vectors ξ:

$$\xi = (\xi_1, \xi_2, \xi_{3,}\xi_4)$$

Where:

$$\xi_i = \sqrt[4]{\prod_{j=1}^{4} u_{ij}} \quad , \quad i = (1, 2, 3, 4)$$

Based on the matrix **P** above, the matrix $\boldsymbol{\xi}$ can be calculated accordingly. Consequently, The weighting matrix w is then obtained by normalizing the characteristic vectors, calculated as follows:

$$w = (w_1 = \frac{\xi_1}{\Sigma\xi_i}, w_2 = \frac{\xi_2}{\Sigma\xi_i}, w_3 = \frac{\xi_3}{\Sigma\xi_i}, w_4 = \frac{\xi_4}{\Sigma\xi_i},)$$

Finally, the comprehensive comfort framework is expressed as

$$C = w_1 T + w_2 V + w_3 N + w_4 Q$$

Where C denotes comprehensive comfort, T denotes thermal comfort, V denotes visual comfort, N denotes acoustic comfort, and Q represents air quality. Hence, the coefficients α_1, α_2, α_3, α_4 are the corresponding weight factors for each comfort aspect.

$$\alpha_1 = w_1 \text{ , } \alpha_2 = w_2 \text{ , } \alpha_3 = w_3 \text{ ,} \alpha_4 = w_4$$

The scale of the different aspects of comfort is not constant. Therefore, in the equation T, V, N, and Q represent the time fraction of comfort hours within a given period. Specifically, T denotes thermal comfort, V denotes visual comfort, A denotes acoustic comfort, and Q represents air quality as before.

Thus, the matrix C can be expressed as

$$C = (w_1\; w_2\; w_3\; w_4) \cdot \begin{pmatrix} T \\ V \\ N \\ Q \end{pmatrix} = w * F$$

This format allows for a comprehensive evaluation of comfort factors across different dimensions and time periods.

According to the regulations outlined in EN16798/EN15251, comfort can be classified into four levels based on the application of the categories used was shown in Table 4.9. These levels provide a structured framework for assessing and categorizing the comfort conditions in buildings. Considering the possible applicability of the categories used, F represents the time fraction matrix of comfort over a specified time period. The time fraction matrix can be expressed as:

$$F = \begin{bmatrix} T_1\, T_2\, T_3\, T_4 \\ V_1\, V_2\, V_3\, V_4 \\ N_1 N_2\, N_3\, N_4 \\ Q_1\, Q_2\, Q_3\, Q_4 \end{bmatrix}$$

where 1, 2, 3, 4 correspond to different categories (Category I, II, III, and IV as classified in Table 4.7). According to the comfort index associated with different comfort types (thermal, visual, acoustic, and air quality), these values follow general quantitative standards as specified in EN16798 regulations.

4.6.2 An Example of Weight Factors Calculation

Here is an example of weight factors calculation based on a client's customised comfort preference.

Consider a client who has provided a customized ranking for the comfort aspects as follows:

Acoustic Comfort > Thermal Comfort > Air Quality > Visual Comfort

The calculation of the weight factors (α_1, α_2, α_3, α_4) is based on the influence matrix P, which represents the impact of each comfort aspect. The matrix is populated according to the client's priorities and the specific influence each factor has.

- Step 1: Define the Influence Matrix P according to Table 4.9.

The influence matrix P is defined as:

$$P = \begin{bmatrix} 0.8\ 0.6\ 0.4\ 0.3 \\ 0.7\ 0.5\ 0.3\ 0.2 \\ 0.6\ 0.4\ 0.2\ 0.1 \\ 0.5\ 0.3\ 0.2\ 0.1 \end{bmatrix}$$

where, the first/second /third /fourth row corresponds to Acoustic Comfort/Thermal Comfort/Air Quality/Visual Comfort.

- Step 2: Calculate the Characteristic Vector ξ

The characteristic vector ξ is calculated using the formula:

$$\xi_i = \sqrt[4]{\prod_{j=1}^{4} u_{ij}}$$

Where ξ_i is the characteristic value for each comfort aspect.

1. Acoustic Comfort $\xi_1 = \sqrt[4]{0.8 \times 0.6 \times 0.3 \times 0.2}$
2. Thermal Comfort $\xi_2 = \sqrt[4]{0.7 \times 0.5 \times 0.3 \times 0.2}$
3. Air Quality $\xi_3 = \sqrt[4]{0.6 \times 0.4 \times 0.2 \times 0.1}$
4. Visual Comfort $\xi_4 = \sqrt[4]{0.5 \times 0.3 \times 0.2 \times 0.1}$

These calculations yield the characteristic vector:

$$\xi = (0.6420, 0.4986, 0.3447, 0.3069)$$

- Step 3: Normalize the Characteristic Vector to Obtain Weight Factors

The weight factors (α_1, α_2, α_3, α_4) are derived by normalizing the characteristic vector:

$$\alpha_{\mathrm{i}} = \frac{\xi_{\mathrm{i}}}{\sum_{i=1}^{4} \xi_i}$$

Calculations:

$$\alpha_3 = \frac{0.3447}{0.6420 + 0.4986 + 0.3447 + 0.3069} = 0.1924$$

$$\alpha_4 = \frac{0.3069}{0.6420 + 0.4986 + 0.3447 + 0.3069} = 0.1711$$

Step 4: Final Results

The final weight factors based on the client's customized preference are:

α_1 (Acoustic Comfort): 0.3582

α_2 (Thermal Comfort): 0.2783

α_3 (Air Quality): 0.1924

α_4 (Visual Comfort): 0.1711

These weight factors reflect the client's customized priorities, providing a quantitative basis for evaluating and optimizing building comfort according to the specified preferences.

5 Apply Machine Learning in the Comprehensive Comfort Framework

5.1 Introduction

With advancements in artificial intelligence, machine learning has addressed the limitations of traditional statistical computing, particularly in the context of big data challenges, and has expanded its application across various fields. As a subfield of artificial intelligence, machine learning focuses on enabling computers to learn and act like humans by autonomously improving their performance through the provision of data and information derived from observations and real-world interactions.

In Chapter 4, the concept of Comprehensive Building Comfort Framework is introduced to emphasize the critical consideration of occupant comfort during the building design phase. Traditionally, comfort assessment methods have relied on various software tools and instrument-based standards, developed through the accumulated expertise of professionals in the field. However, these conventional methods present notable limitations: the lack of integration among software interfaces restricts efficient information exchange, and the reliance on instrument standards often overlooks the human-centred approach that is fundamental to building design.

Chapter 5 builds upon this framework by utilizing machine learning to leverage data from BIM models for the prediction of energy indexes during the building design stage, drawing on the mathematical principles outlined in Chapter 4. This study proposes the application of artificial intelligence (AI) and machine learning (ML) to provide a more integrated, efficient, and intelligent framework for evaluating overall building comfort. The significant volume of data generated by Building Information Modelling (BIM) during the design phase serves as a crucial resource for this analysis. Due to the interconnected and correlated nature of this data, ML algorithms are well-suited to accurately process and analyse these extensive datasets, surpassing the limitations of manual methods. This approach effectively addresses the challenges associated with large-scale data management and enhances the precision of comfort optimization, ensuring that the design remains aligned with human-centred principles.

Additionally, a knowledge-based reasoning engine is developed to assess building comfort, providing a basis for informed decision-making will be present in Chapter 6 later. This engine constructs an expert knowledge base to diagnose building comfort at the design stage, utilizing this knowledge for reasoning and decision support. For AI to effectively diagnose building comfort, it is essential to establish a comprehensive database of building comfort assessments, which can then be used to evaluate and make judgments about new building models and designs before construction begins.

5.2 Application of Machine Learning in Practice

In practice, machine learning involves not just building and training models but also preprocessing and cleaning data, selecting the appropriate model, tuning its parameters, and evaluating its performance. As such, it's a highly interdisciplinary field, requiring knowledge and skills from areas like mathematics, statistics, computer science, and domain-specific knowledge.

The are many advantages of machine learning in daily life. The machine learning can recognize the pattern hide inside of the data and it can help people make better decision and prediction. This algorithm has been used in many aspects, such as medical field, finance area, energy part etc. it give people a higher level of convenient in life than before. The media sites use machine learning technique to sift millions of movies and songs rely on the history preference. Also, the retailers use machine learning to recommend advertise of products according to their customers purchasing behaviour.

There are also many applications of machine learning have been used in the special area of real world. The increasing volume of big data giving machine learning more chance of solving problems in many areas, for instance: Computational finance, for credit scoring and algorithmic trading, Image processing and computer vision, for face recognition, motion detection, and object detection, Computational biology, for tumour detection, drug discovery, and DNA sequencing, Energy production, for price and load forecasting, Automotive, aerospace, and manufacturing, for predictive maintenance, Natural language processing.

5.3 ML Algorithms of Interest and Their Underlying Mathematics

5.3.1 Linear Basis Function Models

Linear Basis function models is one of the simplest models used for regression analysis. It involves a linear combination of input variables and is expressed as:

$$y\,(\mathbf{x}, \mathbf{w}) = w_0 + w_1x_1 + . . . + w_Dx_D$$

where $\mathbf{x} = (x_1, . . ., x_D)^T$. This is often simply known as *linear regression*. The key property of this model is that it is a linear function of the parameters $w_0, . . ., w_D$. It is also, however, a linear function of the input variables x_i, and this imposes significant limitations on the model.

5.3.2 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) are among the most widely adopted artificial intelligence techniques within the domain of building energy management (Bilal et al., 2016). These networks are designed to model complex, non-linear relationships, making them particularly effective in solving multifaceted problems inherent in energy management. ANNs have been successfully applied in various energy-related domains, including the optimization of solar water heating systems, the analysis of solar radiation patterns, electricity consumption monitoring, airflow distribution modeling, energy consumption forecasting, indoor air temperature regulation, and the evaluation and control of HVAC systems (Bilal et al., 2016). The adaptability and versatility of ANNs in these scenarios highlight their utility in optimizing building energy systems and improving overall energy efficiency.

A critical aspect of ANN performance is the quality and diversity of input data. To achieve accurate and reliable predictions, data for ANNs is typically collected from four primary sources: onsite measurements, surveys, billing records, and simulation models (Zhao, 2014). The data collected can encompass a wide range of parameters, such as temperature, humidity, occupancy levels, energy consumption patterns, and equipment operation status. However, raw data often contains noise and inconsistencies, making it essential to apply advanced data pre-processing techniques, such as normalization, filtering, and outlier detection, to refine the data quality before it is fed into the network.

Artificial Neural Networks (ANNs) are computational models inspired by the structure and function of the human brain, designed to identify patterns and model complex relationships in data. The mathematical foundation of ANNs is crucial to understanding how they function, how they are trained, and why they are effective in modelling non-linear and multidimensional phenomena. The key mathematical components and processes of ANNs are outlined as follows.

5.3.2.1 Neural Network Structure of ANNs

An ANN consists of multiple layers of neurons (nodes), typically organized into:

- **Input Layer:** Receives input features (e.g., temperature, humidity, energy usage).
- **Hidden Layers:** One or more intermediate layers where the data is transformed through weighted connections.
- **Output Layer:** Produces the final output (e.g., energy consumption forecast).

Each neuron in a layer receives inputs, processes them, and passes the result to the next layer. The processing involves an activation function, weights, and biases, which are adjusted during training.

5.3.2.2 Mathematical Representation of ANNs

The output of a neuron j in a given layer can be mathematically expressed as:

$$z_j = \sum_{i=1}^{n} w_{ij}\, x_i + b_j$$

where:

x_i are the input values from the previous layer.

w_{ij} are the weights associated with each input connection.

b_j is the bias term.

z_j is the weighted sum output before applying the activation function.

This weighted sum is then passed through an activation function $\sigma(z_j)$ to introduce non-linearity:

$$a_j \ = \sigma(z_j)$$

where a_j is the activated output of the neuron. Common activation functions include:

- Sigmoid: $\sigma(z_j) = \frac{1}{1+e^{-z}}$
- ReLU (Rectified Linear Unit): $\sigma(z_j)$=max (0, z)
- Tanh: $\sigma(z_j) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$

The choice of activation function impacts the model's ability to capture non-linear relationships.

5.3.2.3 Forward Propagation of ANNs

In forward propagation, the input values are passed through the network layer by layer. Each neuron computes a weighted sum of its inputs, applies the activation function, and passes its output to the next layer. This process continues until the output layer is reached, producing the network's prediction y_p.

5.3.2.4 Loss Function

The accuracy of the network's prediction is evaluated using a loss function L(y, y_b), where y is the true value and y_b is the predicted value. Common loss functions include:

- Mean Squared Error (MSE) for regression problems:

$$L(y, y_b) = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_{bi})$$

- Cross-Entropy Loss for classification tasks:

$$L(y, y_b) = -\sum_{i=1}^{N}[y_i \log(y_{bi}) - (1 - y_i)\log(1 - y_{bi})]$$

The choice of loss function depends on the nature of the problem (e.g., regression or classification).

5.3.2.5 Evaluation Metrics

To assess the performance of ANNs, several evaluation metrics are used, such as:

- Accuracy: The proportion of correctly predicted samples.
- Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for regression tasks.

By evaluating these metrics, the effectiveness of the ANN model can be validated, and the model can be fine-tuned accordingly.

5.3.2.6 Summary of ANNs

The underlying mathematics of ANNs is a complex interplay of linear algebra, calculus, and optimization techniques. The success of ANNs in building energy management and other fields lies in their ability to model intricate patterns through a structured learning process that iteratively adjusts weights and biases. By leveraging large datasets and the computational power of modern hardware, ANNs can achieve high accuracy in predicting outcomes and optimizing systems, making them indispensable tools in the evolving field of artificial intelligence.

5.3.3 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) have been extensively applied in building energy analysis over the past decade, particularly in addressing non-linear problems associated with energy consumption prediction. SVMs are well-suited for scenarios with smaller datasets, as they perform efficiently even when limited training data is available. Previous studies demonstrate that SVMs can deliver accurate hourly and monthly predictions of building energy usage. However, despite their effectiveness, SVMs also present several limitations, particularly in scaling to larger datasets and handling complex building scenarios.

One notable example is Zhao et al. (2010), who applied SVMs to predict heating loads in multiple buildings. Although their model produced accurate predictions, the training process was extremely slow due to the large data size involved (Zhao & Magoulès, 2010). To overcome this limitation, Zhao and Magoules (2011) later developed a parallel SVM algorithm to accelerate the training process, demonstrating improvements in speed and efficiency. Li et al. (2010) also explored SVMs alongside general neural networks to predict electricity consumption in buildings, further validating the utility of SVMs in the energy domain.

5.3.3.1 Underlying Mathematics of SVMs

SVMs are fundamentally based on the concept of finding the optimal hyperplane that separates data points belonging to different classes. In the context of regression problems, SVMs aim to find a hyperplane that best fits the data within a specified margin of tolerance, making them particularly effective for non-linear regression tasks like energy prediction.

In the simplest form, for linearly separable data, SVMs seek to identify a hyperplane defined as:

$$f(x) = w^T x + b$$

Where:

w is the weight vector perpendicular to the hyperplane.

x is the input vector.

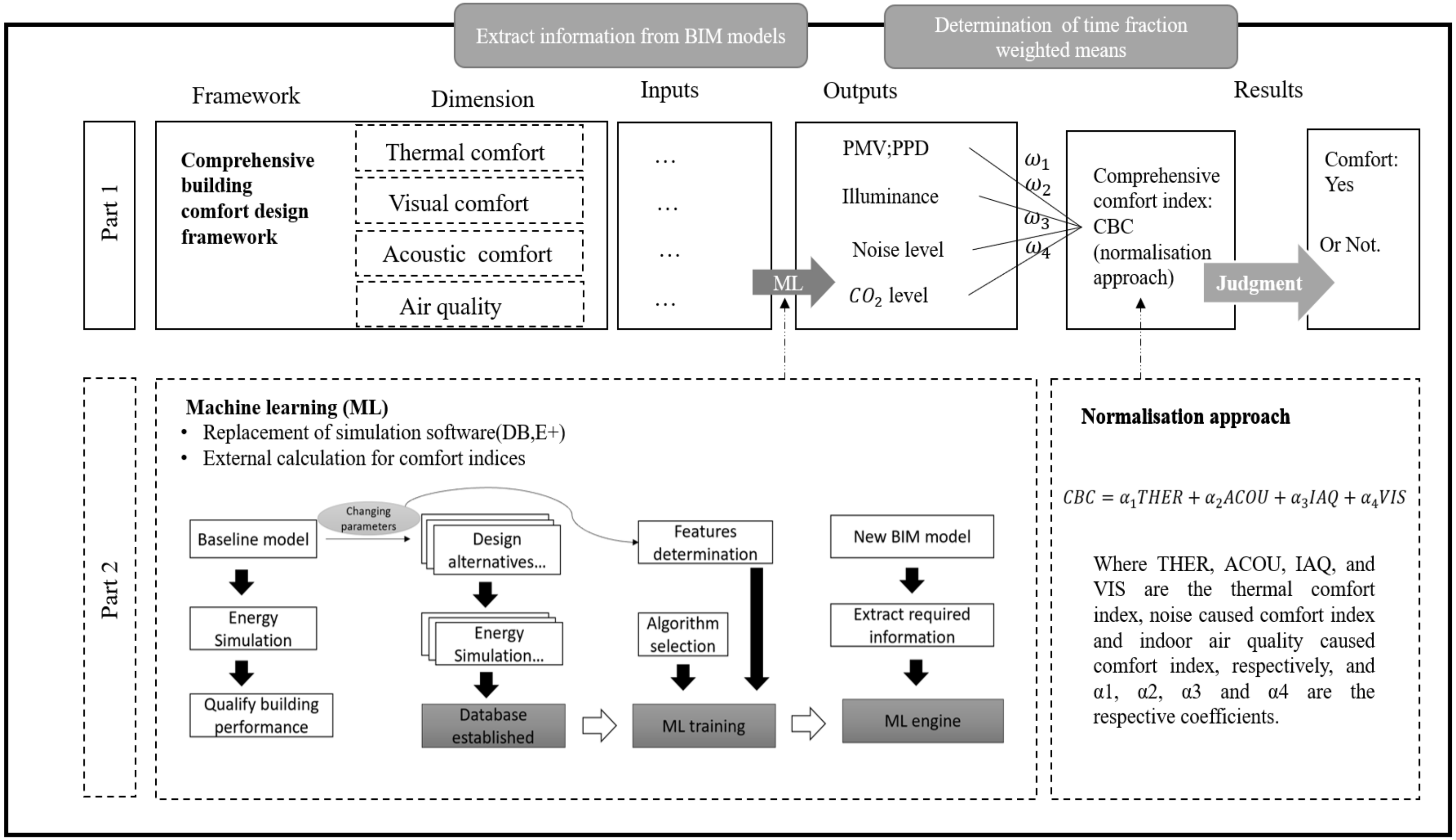
b is the bias term.

The objective is to maximize the margin between the data points and the hyperplane, which enhances the model's ability to generalise. This margin is defined as the distance between the hyperplane and the nearest data points, called support vectors.

5.3.3.2 Evaluation and Performance Metrics

The effectiveness of SVMs in building energy prediction is typically evaluated using metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). These metrics help quantify the accuracy and reliability of SVM models in predicting energy use across different time frames (hourly, daily, monthly

Comprehensive Building Comfort design framework



Extract information from BIM models
Determination of time fraction weighted means
Framework
Dimension
Inputs
Outputs
Results
Part 1
Comprehensive building comfort design framework
Thermal comfort
Visual comfort
Acoustic comfort
Air quality
…
…
…
…
ML
PMV;PPD
Illuminance
Noise level
CO_2 level
ω_1
ω_2
ω_3
ω_4
Comprehensive comfort index: CBC (normalisation approach)
Judgment
Comfort: Yes
Or Not.
Part 2
Machine learning (ML)
Replacement of simulation software(DB,E+)
External calculation for comfort indices
Baseline model
Changing parameters
Design alternatives…
Energy Simulation
Energy Simulation…
Qualify building performance
Database established
Features determination
Algorithm selection
ML training
New BIM model
Extract required information
ML engine
Normalisation approach
$CBC = \alpha_1 THER + \alpha_2 ACOU + \alpha_3 IAQ + \alpha_4 VIS$
Where THER, ACOU, IAQ, and VIS are the thermal comfort index, noise caused comfort index and indoor air quality caused comfort index, respectively, and α1, α2, α3 and α4 are the respective coefficients.

Figure 5. 1 Partially Displayed: The Comprehensive Building Comfort Framework

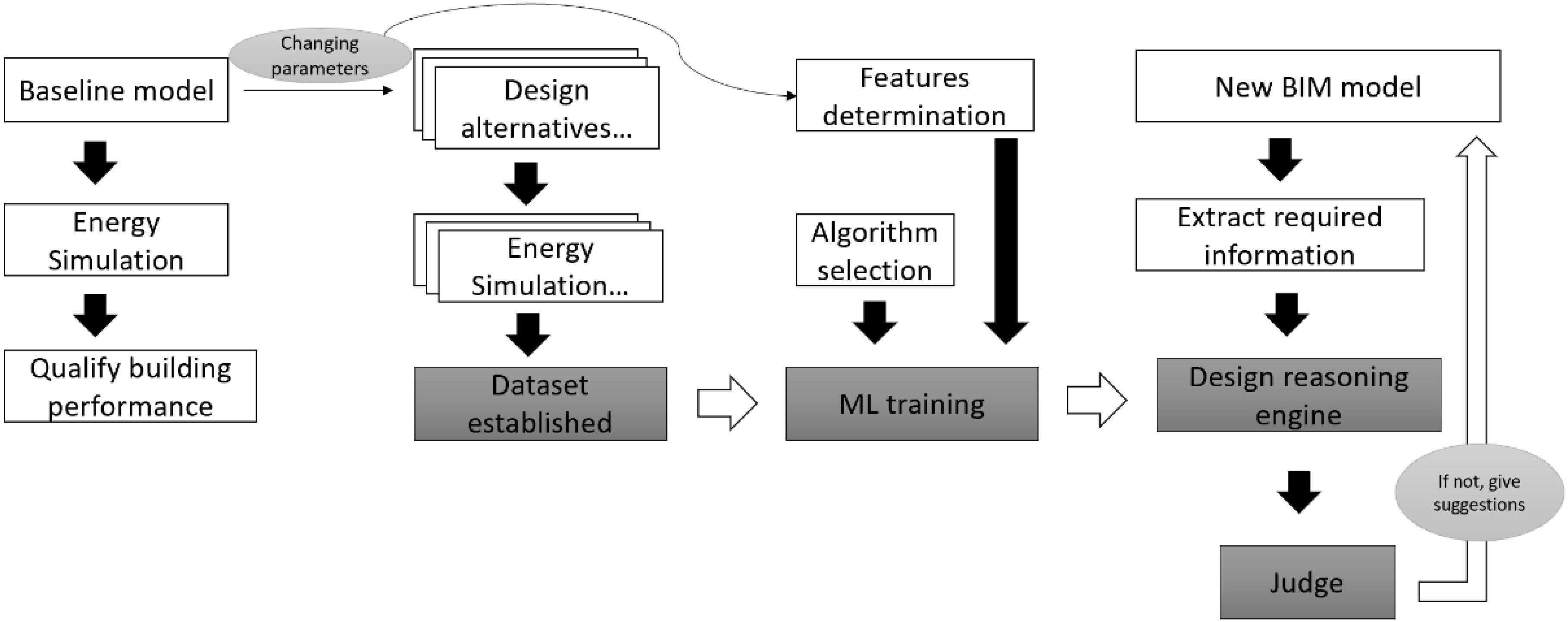


Figure 5. 2 Machine Learning Training Workflow

5.4 The Procedure of ML Training

According to the workflow of Machine Learning (ML) calculations in Figure 5.2, several key steps are involved in this process.

5.4.1 Data Preparation

Data preparation is a critical step in the machine learning process, as the quality and structure of the data directly affect the model's performance.

5.4.1.1 Data Generation

The first step in realising the comprehensive framework is data generation. Data is crucial for developing models in statistical analysis. There are three primary methods of data collection: surveys or questionnaires from occupants, direct measurements in real buildings, and simulations using well-developed software programs. Surveys and measurements, while valuable, are time-consuming and carry a risk of inaccuracies due to environmental and practical constraints. In this research, mock-up data will be utilised to simulate and represent patterns typical of the early design stage, providing a foundation for subsequent analysis. This approach aligns with the analysis of the comprehensive comfort framework discussed in Chapter 4. The second part is data processing, which includes techniques such as data normalization, standardization, and feature engineering. According to the analysis in Chapter 4, the feature scaling of data will be made a decision and prepared for later use. Thirdly, the data splitting part is also important for ensuring the accuracy of the ML training model. In this study, 70% of data were randomly selected for training, while the other 30% will be used in testing.

5.4.1.2 Simulation Tool: Energy Plus

Energy Plus is a well-known energy simulation tool maintained by US Department of energy. Figure 5.3 illustrates the operational framework of the EnergyPlus simulation tool, which is widely employed for building energy performance analysis. The process initiates with the Building Description, representing a detailed input of the building's architectural, structural, and operational characteristics. This description encompasses a wide range of parameters, including material properties, geometric configuration, HVAC systems, and occupant behaviour.

The EnergyPlus Simulation Manager serves as the central computational engine, orchestrating various specialised simulation modules (Zhao, 2014). These modules are integral components responsible for simulating different aspects of building performance:

- The Sky Model Module simulates sky conditions, incorporating factors such as solar radiation and its interactions with the building envelope.
- The Shading Module calculates the impact of shading devices or neighbouring structures on the building's thermal performance, affecting solar gains and energy loads.

- The Air Loop Module manages the building's air distribution systems, simulating heating, ventilation, and air conditioning (HVAC) processes that directly influence indoor climate control.

These modules operate in a cohesive manner under the control of the simulation manager, ensuring a comprehensive and multi-dimensional evaluation of building energy performance. The final output, represented as Calculation Results, provides a detailed analysis of key metrics such as heating and cooling loads, energy consumption, and system efficiency, all derived from the initial building description. This structured simulation workflow allows for a rigorous assessment of energy performance, enabling informed decision-making during the building design and optimisation phases.

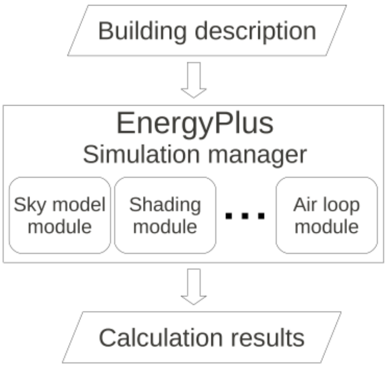


Figure 5. 3 An overview of the simulation tool — EnergyPlus, produced by Zhao (2014).

5.4.2 Machine Learning Training

The second main step is ML training. After data processing, a dataset is established.

5.4.2.1 Feature Selection and Engineering

Before training, the feature determination and algorithm selection will be conducted in this process. Feature selection and engineering are crucial for improving model performance. Use techniques such as correlation analysis, mutual information, or Principal Component Analysis (PCA) to select the most representative and relevant features. Feature engineering involves transforming and generating meaningful features to improve model performance. common techniques such as normalization, standardization, and dimensionality reduction.

5.4.2.2 Model Selection and Training

The model selection process involves evaluating multiple algorithms based on dataset characteristics and analysis requirements. By considering factors such as data size, complexity, and dimensionality, and tuning relevant hyperparameters, the most suitable model is identified for optimal performance.

This section details the process of selecting an appropriate machine learning model, focusing on the data characteristics and specific goals of the analysis. The selection process involves comparing several algorithms, evaluating their strengths and weaknesses, and determining their suitability based on the dataset.

Several common machine learning algorithms are compared, including Decision Trees, Support Vector Machines (SVM), Random Forests, Artificial Neural Networks and so on. The rationale behind choosing these algorithms is based on their effectiveness across different types of predictive tasks:

- Decision Trees (DT): Known for their simplicity and interpretability, decision trees are ideal for small to medium-sized datasets. Overfitting is a risk but can be mitigated through techniques like pruning.
- Support Vector Machines (SVM): SVMs are effective for linear and non-linear classification, particularly with high-dimensional data. They perform well on smaller datasets but may require significant computational resources and careful hyperparameter tuning.
- Random Forests (RF): An ensemble method based on decision trees, Random Forests combine multiple trees to improve accuracy and reduce overfitting, making them suitable for large and complex datasets. However, they may lose interpretability as the number of trees increases.
- ANNs: Powerful models for handling complex, non-linear relationships, particularly effective with large-scale data. These models require extensive computational resources and careful tuning of hyperparameters such as learning rate and layer depth.

There are many model selection criteria. The selection of the appropriate model depends on the characteristics of the dataset, such as the size of the dataset, complexity, and dimensionality. These adjustments enhance the model's accuracy and generalisation ability.

5.4.3 Models Validation and Evaluation

5.4.3.1 Validation methods for ML training

Validation is essential for assessing the performance of a model during the training phase. It helps estimate how well the model generalizes to new, unseen data. Here are some common validation methods:

- Cross-Validation: In this technique, the dataset is divided into k folds. The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, with each fold serving as the test set once. The average performance across all folds gives an estimate of the model's performance. It helps reduce overfitting and ensures the model's robustness.

- K-Fold Validation: A specific type of cross-validation where k represents the number of splits. A common choice is 10-fold cross-validation. It is effective for large datasets as it maximizes training and testing instances, providing a more accurate performance estimate.
- Leave-One-Out Cross-Validation (LOOCV): This is an extreme form of k-fold validation where k equals the number of data points in the dataset. The model is trained on all data points except one, and this process is repeated for each data point. While LOOCV provides an unbiased estimate, it is computationally expensive, especially for large datasets.

5.4.3.2 Evaluation Metrics for Models

To evaluate the effectiveness of the model, various metrics are employed, each offering insight into distinct aspects of the model's performance. The evaluation methods vary depending on whether the model is designed for classification or regression tasks, as each requires specific criteria and metrics tailored to its objectives and output type.

In classification models, metrics such as Accuracy, Precision, Sensitivity (also known as Recall), and F1-Score are used to comprehensively evaluate the model's performance. These metrics are particularly useful for different classification scenarios, including balanced and unbalanced datasets, as each metric provides insight into specific aspects of the model’s accuracy and reliability.

For regression models, metrics like the Correlation Coefficient and Mean Absolute Error (MAE) are applied to assess the model's predictive performance. These metrics offer a quantitative evaluation of how well the regression model predicts continuous values. Below is an explanation of the evaluation metrics commonly used for regression models:

- Correlation Coefficient:

This measures the linear relationship between two variables, specifically the predicted and actual values. It is often used to assess the strength and direction of the linear relationship in regression models. The correlation coefficient (e.g., Pearson's correlation) ranges between -1 and 1:

  - Values close to 1 indicate a strong positive correlation (i.e., the predicted and actual values are closely aligned).
  - Values close to -1 indicate a strong negative correlation.
  - Values close to 0 indicate little to no linear relationship.

It does not directly reflect the magnitude of the prediction error; instead, it shows the linear consistency between the predicted and actual values, making it suitable for evaluating overall trends rather than precise accuracy.

- Mean Absolute Error (MAE):

MAE calculates the average absolute difference between the predicted values and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where y_i is the actual value, and $\hat{y}_i$ is the predicted value.

MAE represents the average prediction error, providing an intuitive and straightforward way to assess the model’s accuracy. A lower MAE indicates more accurate predictions. Unlike the Accuracy metric in Classification models, which is a percentage, MAE provides a concrete measure of error in the same unit as the predicted value, making it highly interpretable.

- Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):

Besides MAE, MSE and RMSE are other common regression metrics that emphasize larger errors by squaring the difference between the predicted and actual values, thus penalizing larger errors more heavily.

In summary, evaluation metrics selection depends on different models. Classification models rely on confusion matrix-based analysis to understand the correct and incorrect classifications, while regression models focus on the magnitude of errors and the linear relationship between predicted and actual values. Precision, Recall, F1-score are ideal for evaluating classification models, which involve discrete categories. Correlation Coefficient, MAE, and similar metrics are suited for evaluating regression models, which involve continuous value predictions. Choosing the right evaluation metrics based on the model’s task type is crucial to accurately assess the model's performance

5.5 Case Study 1 for Single Objective-Daylight Illuminance Prediction

A simple case with limited data is presented in this study to demonstrate the process of extracting energy data from the BIM model to predict real-time daylight illuminance using a machine learning (ML) engine throughout the building's life cycle. The case illustrates the feasibility of utilising weather forecasts to predict future daylight illuminance for an existing building. Consequently, this study proposes a Building Information Modelling (BIM)-based approach for providing data and information to train the relevant ML engine, which is then employed for real-time illuminance prediction.

Based on the reviewed literature and identified research gaps, a prototype integrating a BIM-based machine learning (ML) engine for building energy consumption is proposed, as illustrated in Figure 5.4. BIM files generated by tools such as Revit can be imported into a Graphical User Interface (GUI) depending on the specific file format. The modified files are then mapped onto an energy simulation engine, such as EnergyPlus. Following this, the simulated energy data is fed into the ML engine, which constitutes a critical step in this model. Various ML algorithms can be employed to predict building energy consumption accurately. Furthermore, the outputs from the previous steps can be stored within the smart BIM framework, enabling the management and control of the entire building lifecycle. Consequently, a smart BIM model is developed from the original BIM models, enhancing the overall building management process.

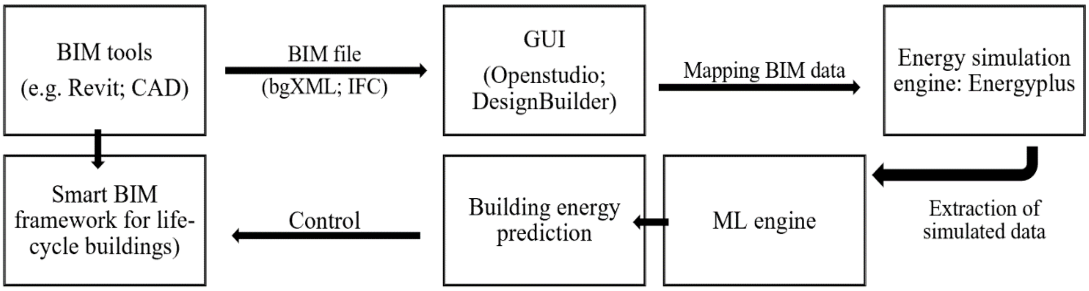


Figure 5. 4 A prototype of a BIM-based ML engine for energy prediction

The target of energy consumption contains 4 main aspects mentioned in Chapter 4, this case study starts from a simple target: daylight illuminance prediction. The whole procedure of machine learning training can be shown in Figure 5.5 below. Designbuilder can easily export the idf file, which can be imported to Energyplus software directly. Energyplus here is being used for dataset simulation. Then, the machine learning algorithms will be used to train the model. Once the machine learning model is trained, a real-time daylight luminance prediction can be made based on the model, given the future weather data for an existing building, which can be used to control the artificial light automatically in a building. Here are the steps below in this case study.

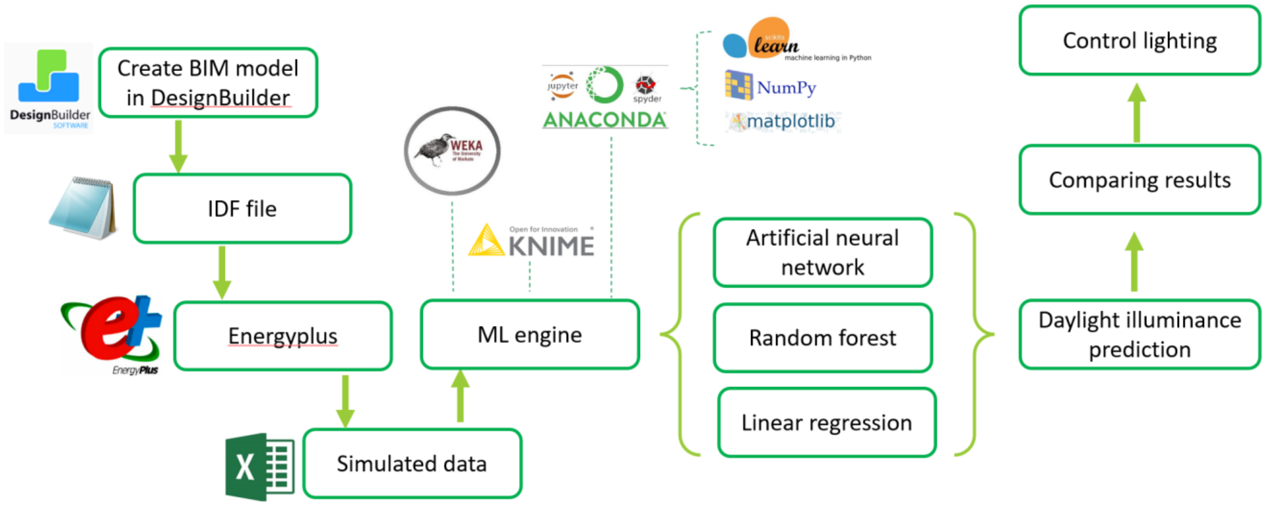


Figure 5. 5 The whole procedure of Machine Learning training for daylight illuminance prediction.

5.5.1 Create BIM Model Using BIM Software

A Building Information Modelling (BIM) model was developed using Designbuilder for a set of three terraced houses situated in Cardiff, as illustrated in Figure 5.6. For the purposes of this study, the middle house in the row has been selected for detailed analysis. The floor plan of this particular house is presented in Figure 5.6. This structure is a two-story, family residence typical of UK housing design. Within this model, the living room located on the ground floor has been chosen specifically for daylight analysis.

To conduct the daylight analysis, a simulated sensor was strategically placed in the centre of the living room, positioned at a height of 0.8 meters above the floor. This height is representative of the typical visual plane for occupants seated in the space, providing an accurate assessment of daylight exposure at eye level. The simulation was carried out over a full annual cycle, spanning from 1 January to 31 December. The location of the house in Cardiff was incorporated into the simulation to account for the region-specific daylight patterns and weather conditions.

Figure 5.7 displays a series of daylight factor (DF) and illuminance (lux) distribution simulations across different floor plans of building spaces. The visualizations are generated using DesignBuilder. It shows how natural light penetrates various rooms or areas under specific daylight conditions. Below is a detailed description of each part in the picture.

1) Simulation result of visual comfort in first floor using DB

This section shows a floor plan with a heat map indicating the daylight factor (DF) and illuminance levels in lux across multiple rooms. The scale on the right ranges from 0 lux (dark areas) to 1278 lux (brightest areas), demonstrating how daylight varies across the space. The distribution reveals that certain rooms receive significantly more light, as shown by the red and orange areas, while others remain in the lower ranges, shown in blue.

2) Simulation result of visual comfort in ground floor using DB.

Similar to the top left, this Figure 5.7 displays ground floor plan with a heat map of daylight distribution. The DF and illuminance values are depicted using a colour gradient from blue (lower values) to red (higher values), and the scale is the same as the previous figure (0 lux to 1278 lux). The pattern here indicates varying light penetration, with brighter zones clustered in specific areas, likely near windows or open spaces.

3) Simulation result of visual comfort in multiple parallel rooms using DB.

This section presents a floor plan with multiple parallel rooms, with daylight distribution indicated in a heat map ranging from 0 lux to 677 lux. The visual shows that some sections of the rooms achieve higher daylight levels, highlighted in red, while the majority of the space remains in the blue spectrum, indicating lower illuminance. This suggests that the daylight is concentrated near specific points, such as windows or openings.

4) Simulation result of visual comfort in living room using DB.

The bottom right quadrant displays a rotated floor plan with a different scale, ranging from 0 lux to 804 lux. The heat map here also illustrates the distribution of daylight, with areas highlighted in blue indicating low light and zones in red and green showing higher light levels. The values marked within the grid cells provide precise measurements of daylight factors, helping to identify how light spreads across this specific space.

This result of visualization provides a comprehensive view of how daylight illuminates various spaces within a building. The colour-coded gradients and corresponding lux scales give insight into the effectiveness of daylight penetration, which is crucial for optimizing natural light use in building design. These visualizations help in assessing the lighting performance of different layouts and assist in making informed decisions to enhance visual comfort and energy efficiency in buildings.

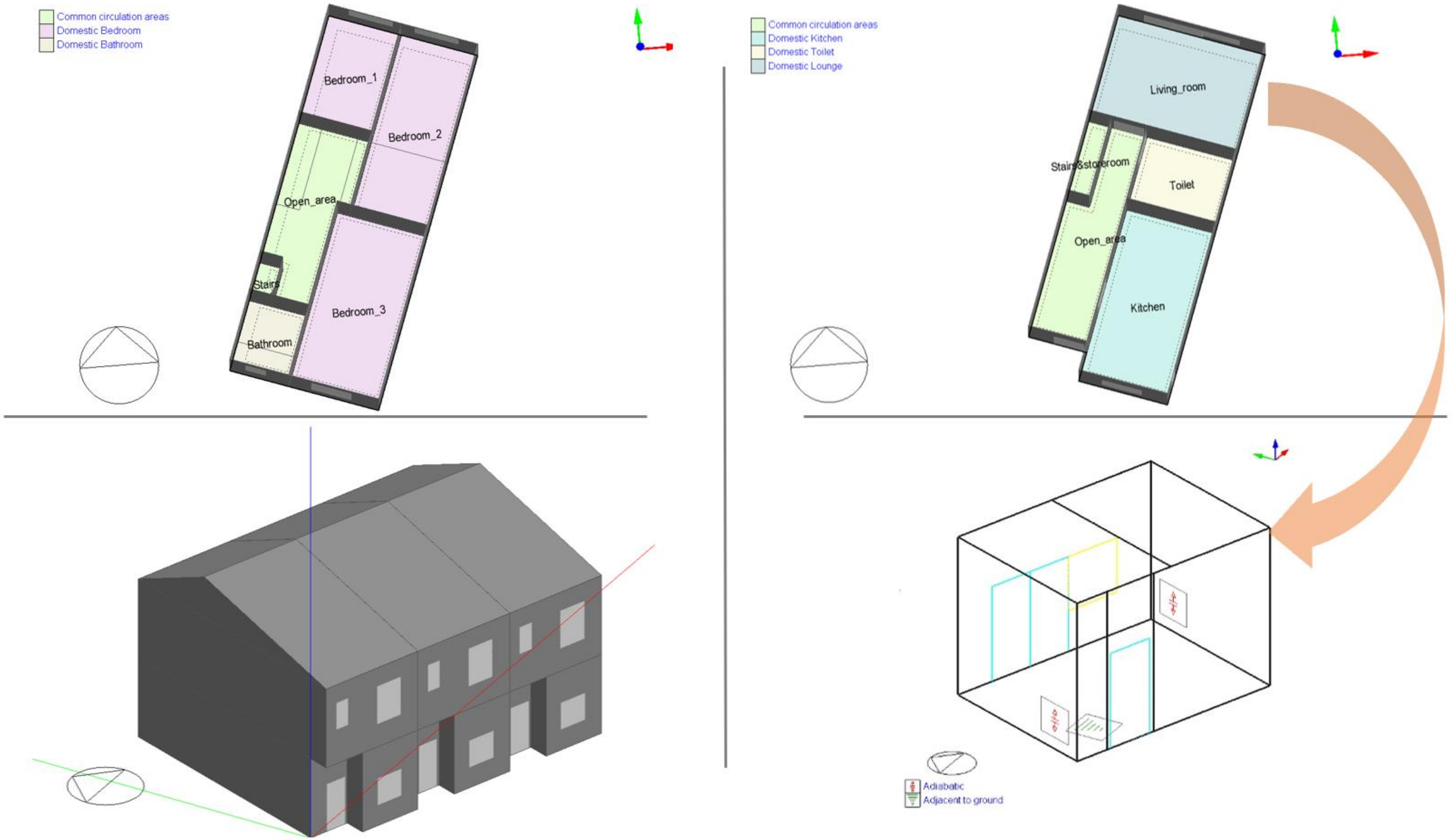


Figure 5. 6 1) A floor plan of the first floor of a single-family house; 2) A floor plan of the ground floor for a single-family house; 3)Three terraced single-family houses model in design builder; 4) The living room in a single-family house.

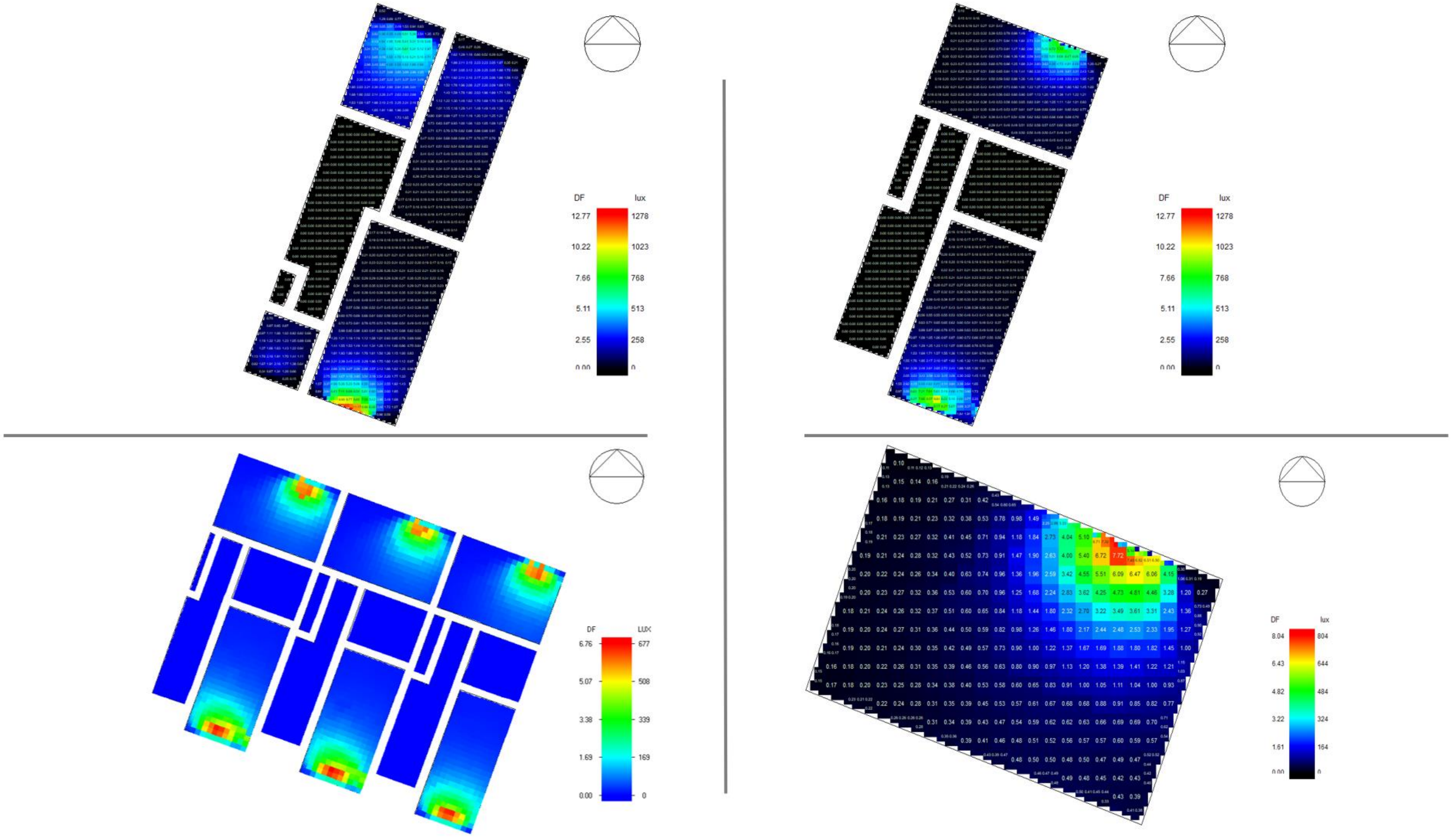


Figure 5. 7 Result of visual comfort simulation in DesignBuilder

5.5.2 Extract Energy Data from a BIM Model

According to the simulation results obtained from DesignBuilder, the data can be manually extracted and organized from the output files. The new dataset is then structured based on the EnergyPlus input and output reference documentation. This dataset comprises ten modifiable variables:

1. Hour of the day
2. Outdoor Air Drybulb Temperature [°C]
3. Site Outdoor Air Humidity Ratio [kgWater/kgDryAir]
4. Site Wind Speed [m/s]
5. Site Diffuse Solar Radiation Rate per Area [W/m²]
6. Site Direct Solar Radiation Rate per Area [W/m²]
7. Site Solar Azimuth Angle [°]
8. Site Solar Altitude Angle [°]
9. Zone Windows Total Transmitted Solar Radiation Rate [W]
10. Daylighting Reference Point Illuminance [lux]

These variables are denoted as X1 to x9 to represent the input variables for the machine learning (ML) engine, while Y signifies the output variable indicating visual comfort (illuminance), which is prepared for ML training (Figure 5.8).

The output dataset generated by EnergyPlus is exported in either .xlsx or .csv format for further analysis in Excel. Figure 5.9 below displays a portion of this output dataset, illustrating how the variables and corresponding data points are organised for ML processing and training. This structured approach ensures the integration of relevant environmental and simulation data, which is critical for predicting visual comfort levels using the ML engine.

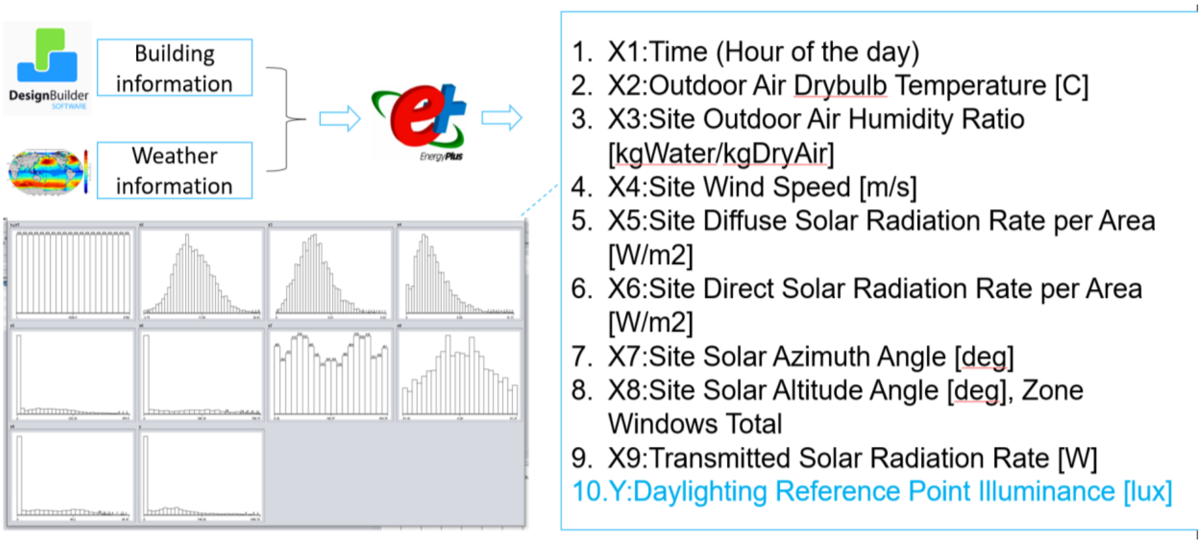
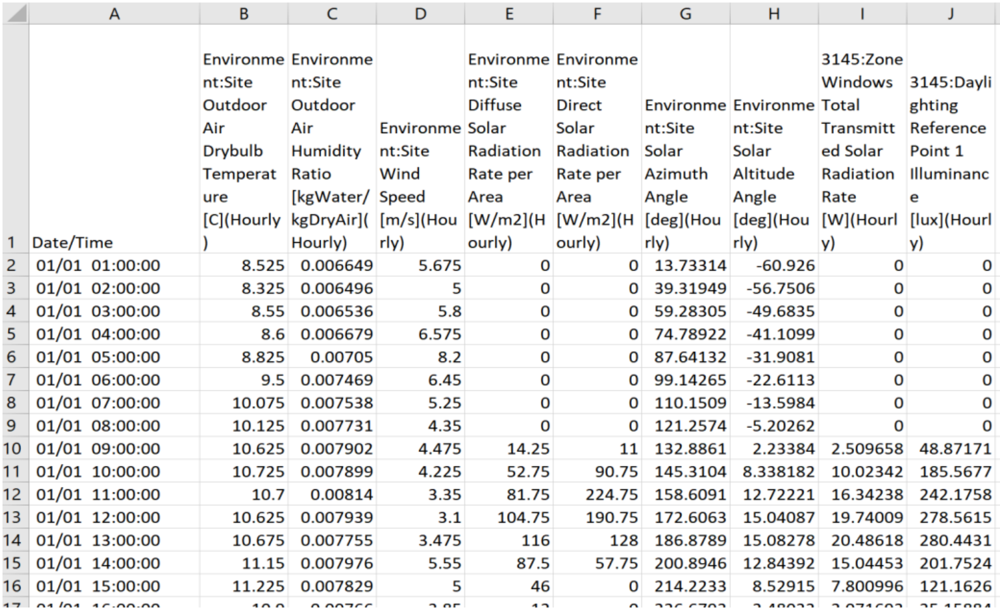

DesignBuilder SOFTWARE
Building information
Weather information
EnergyPlus
1. X1:Time (Hour of the day)
2. X2:Outdoor Air Drybulb Temperature [C]
3. X3:Site Outdoor Air Humidity Ratio [kgWater/kgDryAir]
4. X4:Site Wind Speed [m/s]
5. X5:Site Diffuse Solar Radiation Rate per Area [W/m2]
6. X6:Site Direct Solar Radiation Rate per Area [W/m2]
7. X7:Site Solar Azimuth Angle [deg]
8. X8:Site Solar Altitude Angle [deg], Zone Windows Total
9. X9:Transmitted Solar Radiation Rate [W]
10.Y:Daylighting Reference Point Illuminance [lux]

Figure 5. 8 Data generation and distribution



	A	B	C	D	E	F	G	H	I	J
1	Date/Time	Environment:Site Outdoor Air Drybulb Temperature [C](Hourly)	Environment:Site Outdoor Air Humidity Ratio [kgWater/kgDryAir](Hourly)	Environment:Site Wind Speed [m/s](Hourly)	Environment:Site Diffuse Solar Radiation Rate per Area [W/m2](Hourly)	Environment:Site Direct Solar Radiation Rate per Area [W/m2](Hourly)	Environment:Site Solar Azimuth Angle [deg](Hourly)	Environment:Site Solar Altitude Angle [deg](Hourly)	3145:Zone Windows Total Transmitted Solar Radiation Rate [W](Hourly)	3145:Daylighting Reference Point 1 Illuminance [lux](Hourly)
2	01/01 01:00:00	8.525	0.006649	5.675	0	0	13.73314	-60.926	0	0
3	01/01 02:00:00	8.325	0.006496	5	0	0	39.31949	-56.7506	0	0
4	01/01 03:00:00	8.55	0.006536	5.8	0	0	59.28305	-49.6835	0	0
5	01/01 04:00:00	8.6	0.006679	6.575	0	0	74.78922	-41.1099	0	0
6	01/01 05:00:00	8.825	0.00705	8.2	0	0	87.64132	-31.9081	0	0
7	01/01 06:00:00	9.5	0.007469	6.45	0	0	99.14265	-22.6113	0	0
8	01/01 07:00:00	10.075	0.007538	5.25	0	0	110.1509	-13.5984	0	0
9	01/01 08:00:00	10.125	0.007731	4.35	0	0	121.2574	-5.20262	0	0
10	01/01 09:00:00	10.625	0.007902	4.475	14.25	11	132.8861	2.23384	2.509658	48.87171
11	01/01 10:00:00	10.725	0.007899	4.225	52.75	90.75	145.3104	8.338182	10.02342	185.5677
12	01/01 11:00:00	10.7	0.00814	3.35	81.75	224.75	158.6091	12.72221	16.34238	242.1758
13	01/01 12:00:00	10.625	0.007939	3.1	104.75	190.75	172.6063	15.04087	19.74009	278.5615
14	01/01 13:00:00	10.675	0.007755	3.475	116	128	186.8789	15.08278	20.48618	280.4431
15	01/01 14:00:00	11.15	0.007976	5.55	87.5	57.75	200.8946	12.84392	15.04453	201.7524
16	01/01 15:00:00	11.225	0.007829	5	46	0	214.2233	8.52915	7.800996	121.1626

Figure 5. 9 Partly show the extracted dataset in Excel

5.5.3 ML Algorithms Development

Based on the hourly simulated dataset generated from EnergyPlus, nine input variables have been selected for use in the machine learning (ML) engine, with the sole target variable being daylight illuminance. Figure 5.10 illustrates these nine input variables, and the single output variable utilized within the ML engine. As reviewed in previous sections, the primary ML algorithms applied in this study, such as Artificial Neural Networks (ANN) and Random Forest (RF), require a consistent data format. Therefore, the time step for each record instance is set to hourly, assigning time values from 1 to 8760, corresponding to the total hours in a year. Consequently, the dataset contains 8760 instances.

Weka, a commercial ML software developed using C++, is selected for this task due to its user-friendly interface, particularly for engineers. The dataset, originally in Excel format, can be imported into Weka for normalization and training. In this study, 66% of the dataset is allocated for training, while the remaining 34% is reserved for testing. This partition ensures the model has sufficient data for both training and evaluation, facilitating an effective performance assessment.

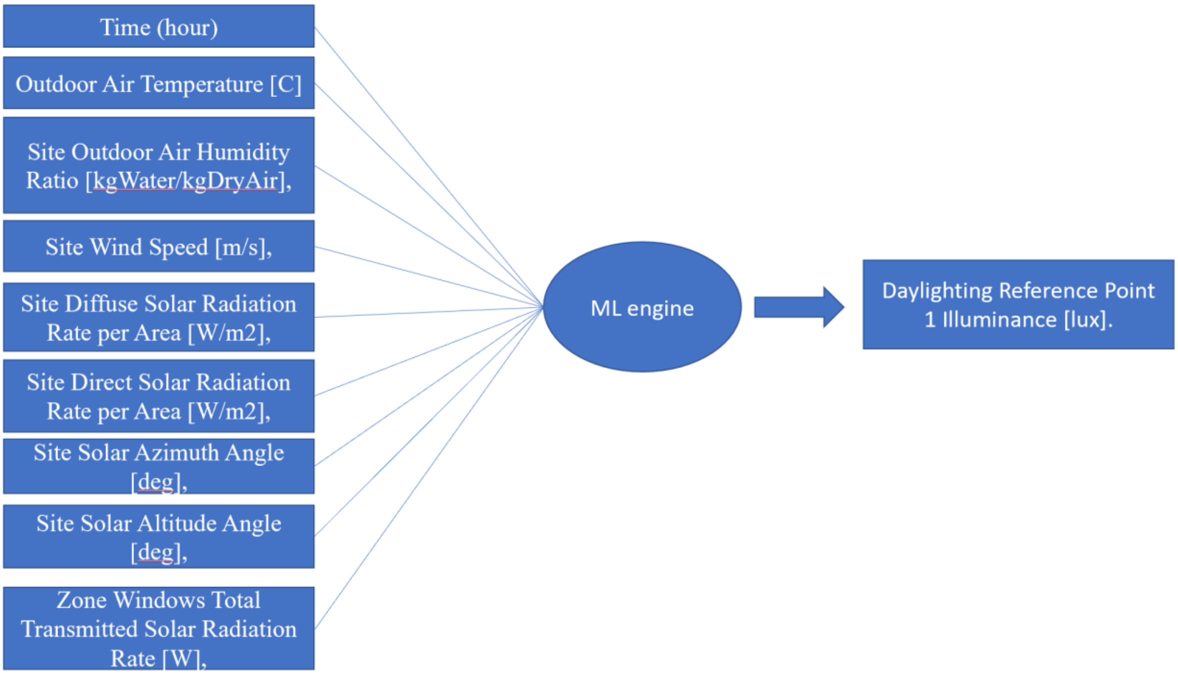

Time (hour)
Outdoor Air Temperature [C]
Site Outdoor Air Humidity Ratio [kgWater/kgDryAir],
Site Wind Speed [m/s],
Site Diffuse Solar Radiation Rate per Area [W/m2],
Site Direct Solar Radiation Rate per Area [W/m2],
Site Solar Azimuth Angle [deg],
Site Solar Altitude Angle [deg],
Zone Windows Total Transmitted Solar Radiation Rate [W],
ML engine
Daylighting Reference Point 1 Illuminance [lux].

Figure 5. 10 The input and output of machine learning engine

5.5.4 ML Algorithm Selected and Comparison in Case Study1 - Linear Regression

The primary machine learning algorithms utilized in this study are Multilayer Perceptron (MLP), Random Forest (RF), and Linear Regression. Each of these algorithms is chosen for its specific strengths in handling various types of data and prediction tasks.

5.5.4.1 Introduction and Basic Theory of Linear Regression

Linear regression is a fundamental statistical modelling technique that represents a continuous response variable as a linear function of one or more predictor variables. Due to its simplicity and ease of interpretation, linear regression is often the first algorithm applied to a new dataset, serving as an initial model to establish baseline performance. It is particularly useful for gaining insights into the relationships between variables and can act as a reference point for evaluating the effectiveness and performance of more complex models, such as neural networks or ensemble methods.

The coefficient of determination, denoted as R^2, is a statistical measure that indicates how well the independent variables (predictors) in a regression model explain the variability of the dependent variable (outcome). It is commonly used to assess the goodness of fit of a linear regression model. Here's how it is calculated:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where,

- ss_{res} is the Residual Sum of Squares, also called the Sum of Squared Errors (SSE). It measures the discrepancy between the observed and predicted values.

Compute the residual sum of squares by summing up the squared differences between each observed value y_i and the predicted value $\hat{y}_i$ from the regression model.

$$ss_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- ss_{tot} is the Total Sum of Squares. It measures the total variance in the observed values of the dependent variable.

Compute the total sum of squares by summing up the squared differences between each observed value y_i and the mean $\bar{y}$

$$ss_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Here, the mean of the observed values $\bar{y}$ can be found from actual dependent variable values y_i

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The interpretation of the result of R^2

- R^2=1, indicates that the regression model perfectly explains the variability of the dependent variable.
- R^2=0, means that the model does not explain any of the variability, and the predictions are as good as using the mean of the observed data.
- 0< R^2 < 1, indicates the proportion of variance explained by the model.

By calculating R^2, the effectiveness of a regression model in capturing the relationship between the variables can be evaluated.

5.5.4.2 The procedure of Employing LR Model

The linear regression model offers a straightforward approach, enabling quick training and testing with minimal computational resources. It also provides clear coefficients that represent the impact of each predictor on the outcome variable, making it highly interpretable and suitable for cases where the focus is on understanding the relationships rather than achieving the highest possible prediction accuracy. Here are the key steps for data processing, training, and visualization of the Linear Regression Model:

Step1 Data processing

- Import CSV File: The dataset is imported as a CSV file to enable structured analysis and manipulation.
- Extract Feature Array X and Target Vector Y: The feature array X represents the predictor variables, while the target vector Y corresponds to the output variable (e.g., daylight illuminance). In this case study, nine feature variables are utilized, as previously mentioned. These variables serve as the inputs for the model, contributing to the prediction of the target variable, which, in this case study, is daylight illuminance. Each feature variable is carefully selected to capture relevant environmental and building parameters that influence the outcome, ensuring the model has sufficient data to learn from and make accurate predictions. The selection of these features is crucial in the development and performance of the machine learning algorithms applied in this study.
- Split the Data into Training and Testing Sets: The dataset is divided into training (70%) and testing (30%) groups to evaluate model performance effectively. The split ensures that the model is trained on a portion of the data and tested on a separate portion to assess its generalizability.

Step2 Training model

- Fitting a linear regression model: A linear regression model is applied to the training dataset, and the model parameters are adjusted to minimize the prediction error.
- Model Score R^2: The coefficient of determination (R^2) for this linear regression model is calculated to be 0.9531, which means that 95.31% of the variance in the dependent variable is explained by the independent variables. It indicates a high level of correlation between the predictor variables and the target variable

In this process, Python was employed for coding, as illustrated in Figure 5.12. The Linear Regression (LR) model demonstrated efficient data processing, producing results almost instantaneously. For comparative analysis, the Support Vector Regression (SVR) algorithm was also applied using the same dataset. Despite the computational efficiency exhibited by LR, the SVR algorithm faced challenges in achieving convergence, likely due to the scattered distribution of the dataset, as evidenced in Figure 5.12.

Step3 Plotting the train model

The visualization of the trained linear regression model is presented in Figure 5.11. This plot demonstrates the relationship between the predictor variables and the predicted output values.

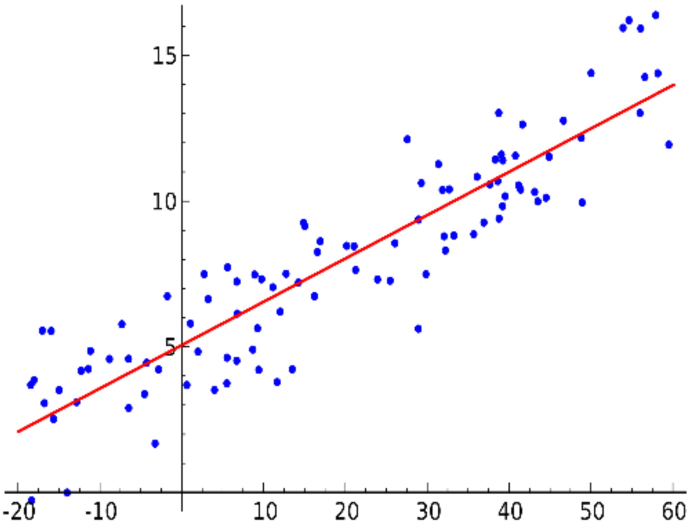


Figure 5. 11 Plotted the training model

5.5.4.3 Results and Findings

The mathematical function representing this relationship can be expressed as follows，: based on the equation derived from the regression model:

$$Y = 0 \times X_1 - 7.3237 \times X_2 + 14583.3888 \times X_3 - 1.8125 \times X_4 - 0.5695 \times X_5 - 0.3652 \times X_6 - 0.1098 \times X_7 + 0.4708 \times X_8 + 18.7163 \times X_9 + 28.0464$$

According to this result, there are some findings for each variable below:

- X_1: Hour of The Day
  - Coefficient: 0
  - The coefficient for "Hour of the day" in this regression model is 0, indicating it has no significant impact on predicting daylight illuminance. Typically, the hour of the day influences daylight intensity due to changes in the sun’s angle and light strength. However, in this model, the zero coefficient for X_1 suggests that the model did not find a meaningful linear relationship between this variable and daylight illuminance.
  - Possible reasons include:
    1) Collinearity: More influential variables, such as solar altitude (X_8)and solar radiation rate (X_9), may have already captured time-related variations. This makes X_1 redundant due to its high correlation with these variables.

2) Data and Feature Engineering: The information in X_1 may have been sufficiently represented by other variables, like solar angles, making its independent contribution less relevant.

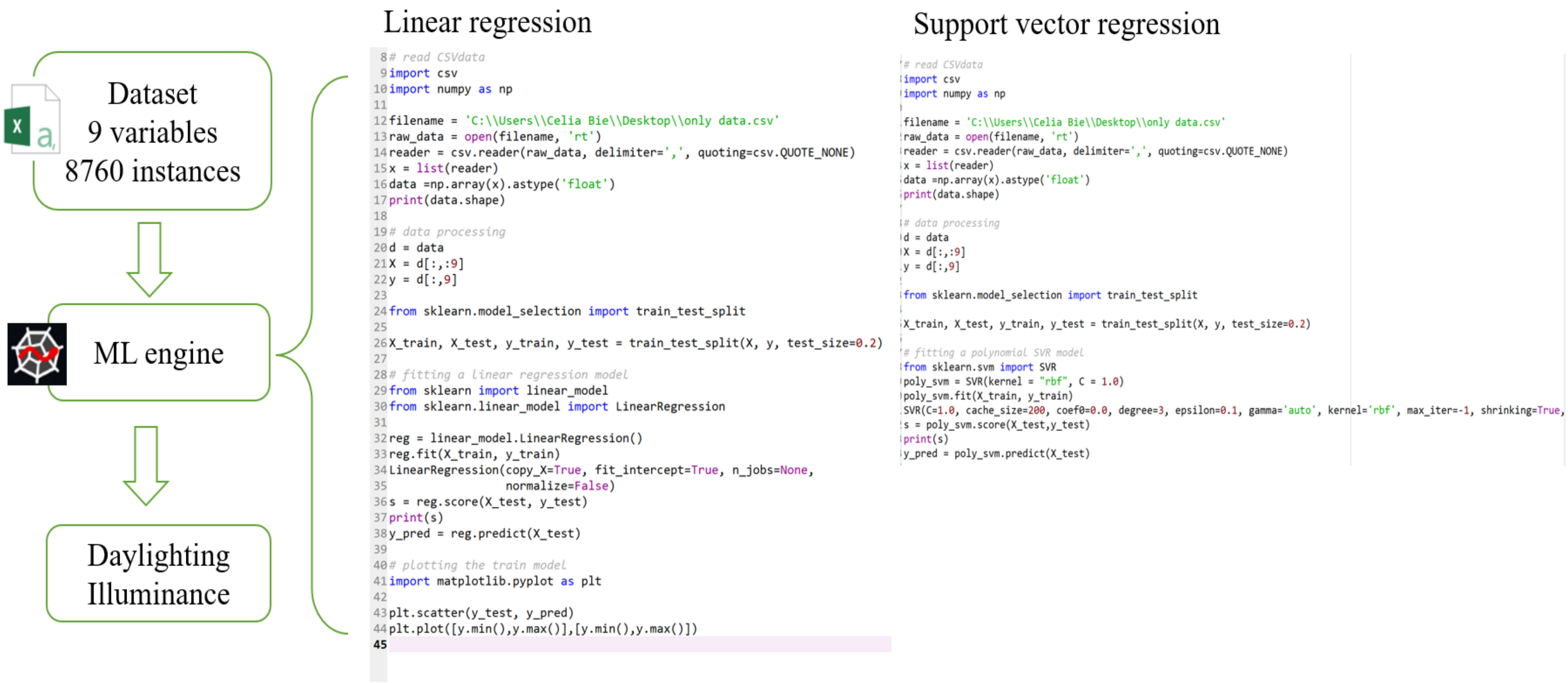
- X_2: Outdoor Air Drybulb Temperature [°C]
  - Coefficient: -7.3237
  - This variable negatively impacts daylight illuminance. For every 1°C increase in dry-bulb temperature, Y decreases by approximately 7.3237 units. This may indicate that higher temperatures could be associated with cloud cover, thereby reducing daylight availability.
- X_3: Site Outdoor Air Humidity Ratio [kgWater/kgDryAir]
  - Coefficient: 14583.3888
  - The humidity ratio shows a substantial positive effect on daylight illuminance, with each unit increase in humidity ratio leading to an increase of approximately 14583.3888 units in Y. This suggests that higher humidity ratios might correlate with clear weather conditions, enhancing daylight levels.
- X_4: Site Wind Speed [m/s]
  - Coefficient: -1.8125
  - Wind speed has a modest negative effect on daylight illuminance. An increase in wind speed may indicate adverse weather conditions, such as storms or high clouds, which could diminish daylight availability.
- X_5: Site Diffuse Solar Radiation Rate per Area [W/m²]
  - Coefficient: -0.5695
  - Diffuse solar radiation has a minor negative impact on Y. This may be due to the fact that higher diffuse radiation often occurs under conditions with cloud cover or atmospheric scattering, reducing direct sunlight.
- X_6: Site Direct Solar Radiation Rate per Area [W/m²]
  - Coefficient: -0.3652
  - Although direct solar radiation typically correlates with increased daylight, the small negative coefficient suggests that other factors, such as cloud cover or structural obstructions, might counteract its influence in this case study.
- X_7: Site Solar Azimuth Angle [°]
  - Coefficient: -0.1098
  - The solar azimuth angle has a negligible negative effect on daylight illuminance, likely reflecting the fact that azimuthal changes do not substantially affect overall light levels unless combined with other conditions, such as time of day.
- X_8: Site Solar Altitude Angle [°]

  - Coefficient: 0.4708
  - The solar altitude angle has a slight positive effect on Y. A higher solar altitude generally corresponds to more direct sunlight reaching the site, thus increasing daylight availability.
- X9**:** Zone Windows Total Transmitted Solar Radiation Rate [W]
  - Coefficient: 18.7163
  - This variable shows a significant positive influence on daylight illuminance. As it directly measures the solar radiation transmitted through windows, it acts as a critical indicator of daylight levels within the building.
- Intercept: 28.0464
  - This term represents the baseline level of daylight illuminance when all other variables are zero. It serves as the initial reference point and may account for other constant sources of light or baseline environmental conditions.

5.5.4.4 Analysis and Conclusion

In summary, the regression model highlights the different contributions of each variable to daylight illuminance within a building environment in this case study. The most influential variable is the outdoor air humidity ratio (X_3), followed closely by the transmitted solar radiation rate through windows (X_9). These two variables significantly enhance the daylight levels, indicating their critical roles in predicting and optimizing building daylight performance. While, the negative coefficients associated with variables such as wind speed, temperature, and diffuse radiation suggest that these factors, when increased, generally correlate with reduced daylight availability, possibly due to weather conditions like cloudiness. Meanwhile, other variables, such as the solar azimuth and altitude angles, show minor effects, reflecting their secondary roles in determining daylight levels.

In this case study, this regression model provides valuable insights into the complex interactions between environmental conditions and daylight illuminance in building contexts. Understanding these relationships is essential for optimizing building design to maximize natural light, improving both energy efficiency and occupant comfort.



Linear regression

```
# read CSVdata
import csv
import numpy as np

filename = 'C:\\Users\\Celia Bie\\Desktop\\only data.csv'
raw_data = open(filename, 'rt')
reader = csv.reader(raw_data, delimiter=',', quoting=csv.QUOTE_NONE)
x = list(reader)
data =np.array(x).astype('float')
print(data.shape)

# data processing
d = data
X = d[:,:9]
y = d[:,9]

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# fitting a linear regression model
from sklearn import linear_model
from sklearn.linear_model import LinearRegression

reg = linear_model.LinearRegression()
reg.fit(X_train, y_train)
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                 normalize=False)
s = reg.score(X_test, y_test)
print(s)
y_pred = reg.predict(X_test)

# plotting the train model
import matplotlib.pyplot as plt

plt.scatter(y_test, y_pred)
plt.plot([y.min(),y.max()],[y.min(),y.max()])
```

Support vector regression

```
# read CSVdata
import csv
import numpy as np

filename = 'C:\\Users\\Celia Bie\\Desktop\\only data.csv'
raw_data = open(filename, 'rt')
reader = csv.reader(raw_data, delimiter=',', quoting=csv.QUOTE_NONE)
x = list(reader)
data =np.array(x).astype('float')
print(data.shape)

# data processing
d = data
X = d[:,:9]
y = d[:,9]

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# fitting a polynomial SVR model
from sklearn.svm import SVR
poly_svm = SVR(kernel = "rbf", C = 1.0)
poly_svm.fit(X_train, y_train)
SVR(C=1.0, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma='auto', kernel='rbf', max_iter=-1, shrinking=True,
s = poly_svm.score(X_test,y_test)
print(s)
y_pred = poly_svm.predict(X_test)
```

Figure 5. 12 The Python script for linear regression and support vector regression, respectively.

5.5.5 ML Algorithms Development- Artificial Neural Network (ANN)

5.5.5.1 Basic Theory and Structure of ANN

An Artificial Neural Network (ANN) consists of highly interconnected layers of neurons that model complex relationships between inputs and desired outputs. The network undergoes training by iteratively adjusting the connection weights, enabling it to map input data to the corresponding outputs accurately. In this case study the multilayer perception technology is used here.

Multilayer Perceptron (MLP): This is a widely used feedforward neural network that involves multiple layers of neurons. It processes information in a forward direction—from input through hidden layers to output—without any feedback loops. The network learns through a process called backpropagation, which adjusts the weights to minimize prediction errors.

Mathematics:

$$y = f(w_{10} + w_{11}x_{1+}w_{12}x_2 + \cdots + w_{1d}x_d) = f(\boldsymbol{W}_1^T\boldsymbol{x} + w_{10})$$

$$f = \begin{cases} 1, & \boldsymbol{W}_1^T\boldsymbol{x} + w_{10} > 0 \\ 0, & O.W. \end{cases}$$

$$\boldsymbol{W}_1 = \begin{bmatrix} w_{11} \\ \vdots \\ w_{1d} \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

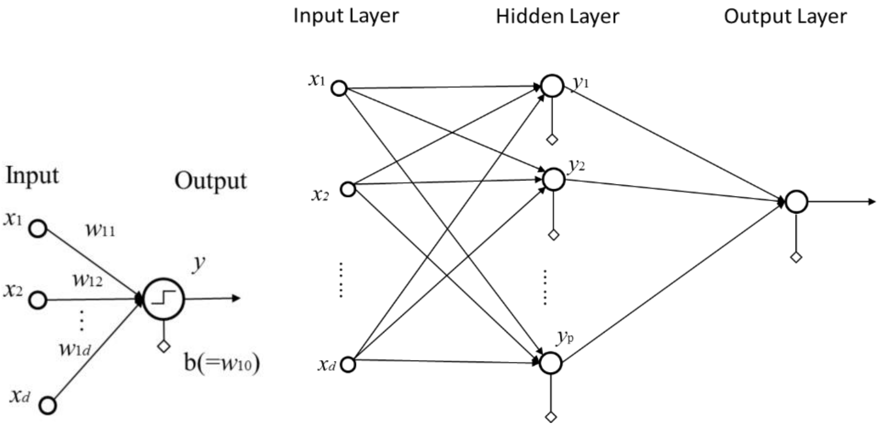


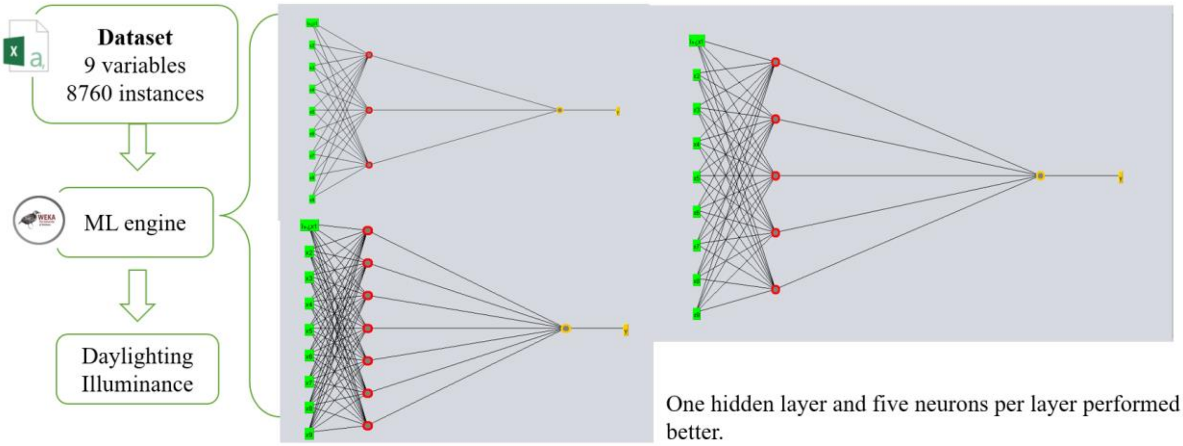
Figure 5. 13 Structure of Multilayer Perceptron

As Figure 5.13 illustrated above, the MLP structure can be summarized as a sequence of linear transformations followed by non-linear activations, where the learning process involves adjusting the weights and biases to minimize the prediction error. The combination of hidden layers and non-linear activations enables the MLP to model complex functions and make accurate predictions.

5.5.5.2 Steps for Implementing ANN

- Data Preprocessing:
  - Input Data: Organize the dataset with 9 input variables (features) such as outdoor temperature, humidity ratio, wind speed, solar radiation, etc, which mentioned before. The target variable y is the daylight illuminance.
  - Normalization: ANN requires the input features to be scaled. Normalize the input data to a range, typically between 0 and 1, or standardize the data (mean 0, variance 1) to ensure the network can train efficiently.
  - Splitting the Data: Split the dataset into training (70%) and testing (30%) sets.
- Define the ANN Architecture:
  - Input Layer: 9 neurons (for the 9 input variables).
  - Hidden Layers: Start with 1 hidden layers with a varying number of neurons. Use ReLU (Rectified Linear Unit) as the activation function.
  - Output Layer: For regression (predicting daylight illuminance), a single neuron is used with a linear activation function.
- Compile the ANN Model:

Use Mean Squared Error (MSE) as the loss function since this is a regression problem. One hidden layer is selected and 3,5 and 7 neurons per layer is chosen respectively (Figure 5.14). After comparing the results from different neurons per layer, we find that the neuron network with 5 neurons per layer performed better than other options here. Hence, the ANN built in this experiment has only one hidden layer and 5 neurons within.



One hidden layer and five neurons per layer performed better.

Figure 5. 14 Step for implementing ANN in case study 1

5.5.6 ML Algorithms Development- Random Forest (RF)

Random forests are an ensemble learning method used for classification, regression, and various other tasks. They operate by constructing a large number of decision trees during training and then producing a final output that is either:

- For classification tasks: The mode of the classes predicted by individual trees.
- For regression tasks: The mean of the predictions from individual trees.

Here are the key Components of RF:

- Decision Trees:
  Each decision tree in the forest is trained on a subset of the data. A decision tree makes predictions by recursively splitting the data based on feature values that best separate the data points into different target classes or values.
- Bagging (Bootstrap Aggregating):
  Bagging is a technique used to train each decision tree on a randomly selected subset of the training data. This random sampling is done with replacement, meaning some data points may appear multiple times in the subset. The idea behind bagging is to reduce variance and improve model generalization by averaging multiple predictions from different models trained on slightly different data.

5.5.6.1 The advantages of RF:

A single decision tree tends to overfit the training data, capturing noise and complex patterns that may not generalize well to unseen data. By averaging the predictions from multiple decision trees trained on different subsets of the data, random forests mitigate the overfitting problem. The variability of individual trees is reduced by aggregating their outputs, leading to better generalization performance.

Here is Random Forest Formula: Random Forests = Decision Trees + Bagging

Bagging introduces randomness in the training process by using different subsets of the training data. Randomness in feature selection: During the training of each tree, a random subset of features is considered at each split, further improving the diversity of the trees and reducing correlation between them.

In conclusion, Random forests leverage the strength of decision trees while addressing their weaknesses, such as overfitting, by using bagging and randomness in feature selection. This makes random forests a powerful and robust method for both classification and regression tasks.

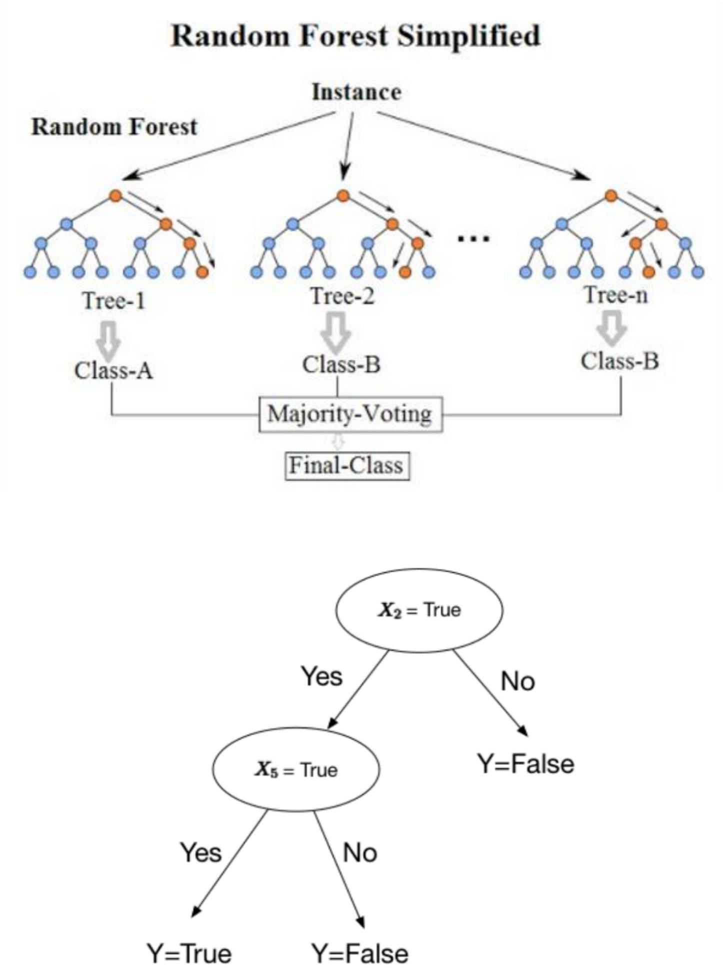


Figure 5. 15 The structure of Random Forest

5.5.6.2 Steps for implementing Random Forest (RF)

Here are some key steps for implementing Random Forest

- Data Preprocessing: Organize the dataset with 9 input variables (features) and the target variable (daylight illuminance).
- Feature Importance: No need to scale the input data for Random Forests since they are not sensitive to the scale of the data.
- Splitting the Data: Similar to the ANN, split the dataset into training (e.g., 70%) and testing (e.g., 30%) sets.
- Train the Random Forest Model: Initialize the Random Forest Regressor with parameters such as the number of trees (n_estimators), maximum depth of the tree, and minimum samples for a leaf.
- Feature Importance: After training, evaluate which input features (variables) contributed most to the prediction.

- Model Evaluation: Evaluate the model on the test set using metrics such as Mean Squared Error (MSE) and R-squared (R^2).
- Prediction: Make predictions using the trained Random Forest model.

5.6 Comparison and Evaluation

Random forest has the high degree of accuracy and the high speed of learning process. Here the bag size percent and batch size are both 100. The tree depth is unlimited here in each tree of the random forest. Linear regression is used for comparation with other two methods here. Cross-validation rank in 10 for detecting the errors. The experiment compares these three methods from five aspects: correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error. The table 5.1 presents the results of three different machine learning algorithms—Multilayer Perceptron, Random Forest, and Linear Regression—based on five evaluation metrics: correlation coefficient, mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and root relative squared error (RRSE). Here's an analysis of the results:

1. Correlation Coefficient:
   Multilayer Perceptron achieved a correlation coefficient of 0.9948, indicating a very strong relationship between the predicted and actual values. Random Forest had the highest correlation coefficient of 0.9991, showing an almost perfect prediction performance. Linear Regression had a correlation coefficient of 0.9531, which is significantly lower than the other two methods, indicating it struggles with capturing the patterns in this dataset.
2. Mean Absolute Error (MAE):
   Multilayer Perceptron resulted in an MAE of 21.6228, meaning the average absolute difference between the predicted and actual values was 21.6 units. Random Forest achieved a much lower MAE of 6.3013, showing that this method has the smallest average error among the three. Linear Regression had an MAE of 52.6389, the highest among the three, which suggests that it performs poorly in terms of minimizing errors in this dataset.
3. Root Mean Squared Error (RMSE): Multilayer Perceptron had an RMSE of 30.5761, indicating a moderate spread in the prediction errors. Random Forest again outperformed with an RMSE of 12.5186, showing better prediction accuracy and less error. Linear Regression had the highest RMSE of 63.1839, which shows that it has the largest error spread among all methods.
4. Relative Absolute Error (RAE): Multilayer Perceptron had an RAE of 9.15%, meaning its prediction errors were 9.15% of the errors a baseline predictor (mean) would make. Random Forest showed the best performance here with an RAE of 2.67%, implying that it performs extremely well relative to a baseline. Linear Regression had a much higher RAE of 22.36%, again suggesting that it performed significantly worse in this metric.

5. Root Relative Squared Error (RRSE):
   Multilayer Perceptron scored 10.51% in RRSE, indicating it reduced the error to about 10.5% compared to a baseline prediction. Random Forest had the best RRSE at 4.30%, further demonstrating its superior accuracy in this dataset. Linear Regression had an RRSE of 30.26%, the highest among the three, indicating the worst performance in this category as well.

Table 5. 1 The Result of Each Algorithms

	Correlation coefficient	Mean absolute error	Root mean squared error $\mathrm{RMSE} = \sqrt{\frac{\sum_{n}^{i=1}(Y_i - \hat{Y}_i)^2}{n}}$	Relative absolute error $\mathrm{MAE} = \frac{\sum_{n}^{i=1}\lvert Y_i - \hat{Y}_i\rvert}{n}$	Root relative squared error	Total Number of Instances
Multilayer Perceptron (1 hidden layer 5 neurons per layer)	0.9948	21.6228	30.5761	9.15%	10.51%	8760
Random Forest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1	0.9991	6.3013	12.5186	2.67%	4.30%	8760
Linear Regression	0.9531	52.6389	88.016	22.28%	30.26%	8760

In summary, the correlation coefficient for all the algorithms presented here is close to 1, indicating a strong relationship between the input variables and the output variables. This suggests that daylight illuminance is highly influenced by factors such as the time of day and weather conditions, with additional factors like surrounding buildings and trees playing a role as well. Random Forest outperforms the other two methods across all evaluation metrics, particularly in MAE, RMSE, RAE, and RRSE. Its near-perfect correlation coefficient of 0.9991 indicates a highly accurate model that effectively captures the relationships in the data. This, combined with its lower error rates, makes

Random Forest the most suitable choice for this dataset. While Multilayer Perceptron also performs well, its error metrics are slightly higher, making it a less precise option compared to Random Forest. On the other hand, Linear Regression demonstrates the weakest performance, with significantly higher error rates and a lower correlation coefficient. This suggests that Linear Regression is not well-suited to handle the complexity of this dataset, where non-linear relationships are better captured by Random Forest and Multilayer Perceptron. Consequently, Random Forest emerges as the most effective algorithm for this case study, followed by Multilayer Perceptron, with Linear Regression being the least appropriate due to its limitations in modelling non-linear patterns. Thus, the prediction of future daylight illuminance through the weather forecast is feasible in building design stage.

5.7 Case study 2 for Prediction of PMV-PPD

5.7.1 Data Generation and Feature Definition.

In this case study, the research model is same as the builds used in case study 1. as the layout of the building is illustrated before in Figure 5.6. The initial room features for the model are selected based on the analysis from Chapter 4. These features represent a combination of environmental conditions and specific room characteristics. The environmental variables capture the external factors that affect indoor conditions, while the room-specific features focus on the structural and occupancy details of the space. The selected features are as follows:

X1. Environment: Site Outdoor Air Drybulb Temperature [°C] (Hourly) – Represents the outside air temperature.

X2. Environment: Site Outdoor Air Dewpoint Temperature [°C] (Hourly) – Indicates the temperature at which air becomes saturated with moisture.

X3. Environment: Site Outdoor Air Barometric Pressure [Pa] (Hourly) – Reflects the atmospheric pressure at the site location.

X4. Environment: Site Wind Speed [m/s] (Hourly) – Measures the speed of the wind at the site.

X5. Environment: Site Wind Direction [°](Hourly) – Specifies the direction from which the wind is blowing.

X6. Environment: Site Diffuse Solar Radiation Rate per Area [W/m^2] (Hourly) – Refers to solar radiation that is scattered by the atmosphere.

X7. Environment: Site Direct Solar Radiation Rate per Area [W/m^2] (Hourly) – Represents the direct beam radiation received on a surface.

X8. Environment: Site Solar Azimuth Angle [°] (Hourly) – Describes the compass direction from which the sunlight is coming.

X9. Environment: Site Solar Altitude Angle [°] (Hourly) – Represents the height of the sun above the horizon.

X10. Volume [m^3] – The total volume of the room.

X11. People [m² per person]– The occupant density in terms of square meters per person.

X12. Ground floor area [m²] – The size of the ground floor of the building.

X13. U-value of the floor [W/K-m²] – Measures the heat transfer rate through the floor.

X14. Ceiling area [m²] – The total ceiling area in the room.

X15. U-value of the ceiling [W/K-m²] – Indicates the thermal transmittance of the ceiling.

X16. Door area [m²] – The size of the doors in the room.

X17. U-value of the door [W/K-m²] – Measures the insulation properties of the door.

X18. Total internal wall area [m²] – The area of the internal walls within the room.

X19. U-value of internal walls [W/K-m²] – Reflects the heat transfer rate through the internal walls.

X20. External wall size [m²] – The area of the external walls.

X21. U-value of external walls [W/K-m²] – Indicates the insulation properties of the external walls.

X22. Glazing area [m²] – The area of the glazed surfaces (windows).

X23. U-value of glazing [W/K-m²] – Measures the thermal transmittance of the glazing.

X24. Window glass thickness [m] – The thickness of the glass used in the windows.

X25. Zone Mean Radiant Temperature [°C](Hourly) – The average temperature of the surfaces surrounding the occupant.

X26. Zone Mean Air Temperature [°C] (Hourly: ON) – The average air temperature in the zone.

X27. Zone Operative Temperature [°C] (Hourly: ON) – A combined measure of air temperature and radiant temperature felt by occupants.

X28. Zone Air Relative Humidity [%] (Hourly: ON) – The amount of moisture in the air as a percentage of the total moisture the air can hold.

There are 28 input parameters totally, represented as X1-X28. The output variables for thermal comfort prediction are as follows:

Y1. Y1. Zone Thermal Comfort Fanger Model PMV (Hourly) – This is the predicted mean vote (PMV) index, which assesses the thermal sensation of occupants.

Y2. Y2. Zone Thermal Comfort Fanger Model PPD [%](Hourly) – This represents the predicted percentage of dissatisfied (PPD) occupants with the thermal conditions.

Figure 5.16 illustrates the data distribution for these variables, revealing an uneven distribution. This non-uniformity in the data may affect the model’s performance and emphasizes the importance of proper feature selection and data handling for accurate prediction of thermal comfort. Figure 5.16 displays the distribution of various input variables used in the model, particularly focusing on U-values for different building components, as labelled with X13, X15, X17, X19, X21, X23, and X24. U-values represent the thermal transmittance of building materials, which are critical for energy performance analysis. These variables are essential for determining the thermal insulation properties of floors, ceilings, doors, internal and external walls, glazing, and window glass thickness. Each histogram shows the range and frequency of values for a specific input variable.

For example:

X13 (U-value of the floor): Shows the distribution of U-values, reflecting the insulation characteristics of floors.

X15 (U-value of the ceiling): Indicates the thermal transmittance for ceiling insulation.

X17 and X19 (U-values of the door and internal walls): Represent the heat transfer through doors and internal walls.

X21 (U-value of external wall): A key variable impacting heat loss or gain through the building’s external envelope.

X23 (U-value of glazing): Describes the insulation performance of windows and glazing.

X24 (Window glass thickness): Reflects the variation in glass thickness, which can influence heat retention and daylighting.

This detailed input parameter dataset is fundamental in optimizing building energy consumption, ensuring the model can accurately predict thermal performance and comfort outcomes. The distribution graphs help identify how these variables vary across different room layout, aiding in the calibration of machine learning models for energy performance and comfort predictions.

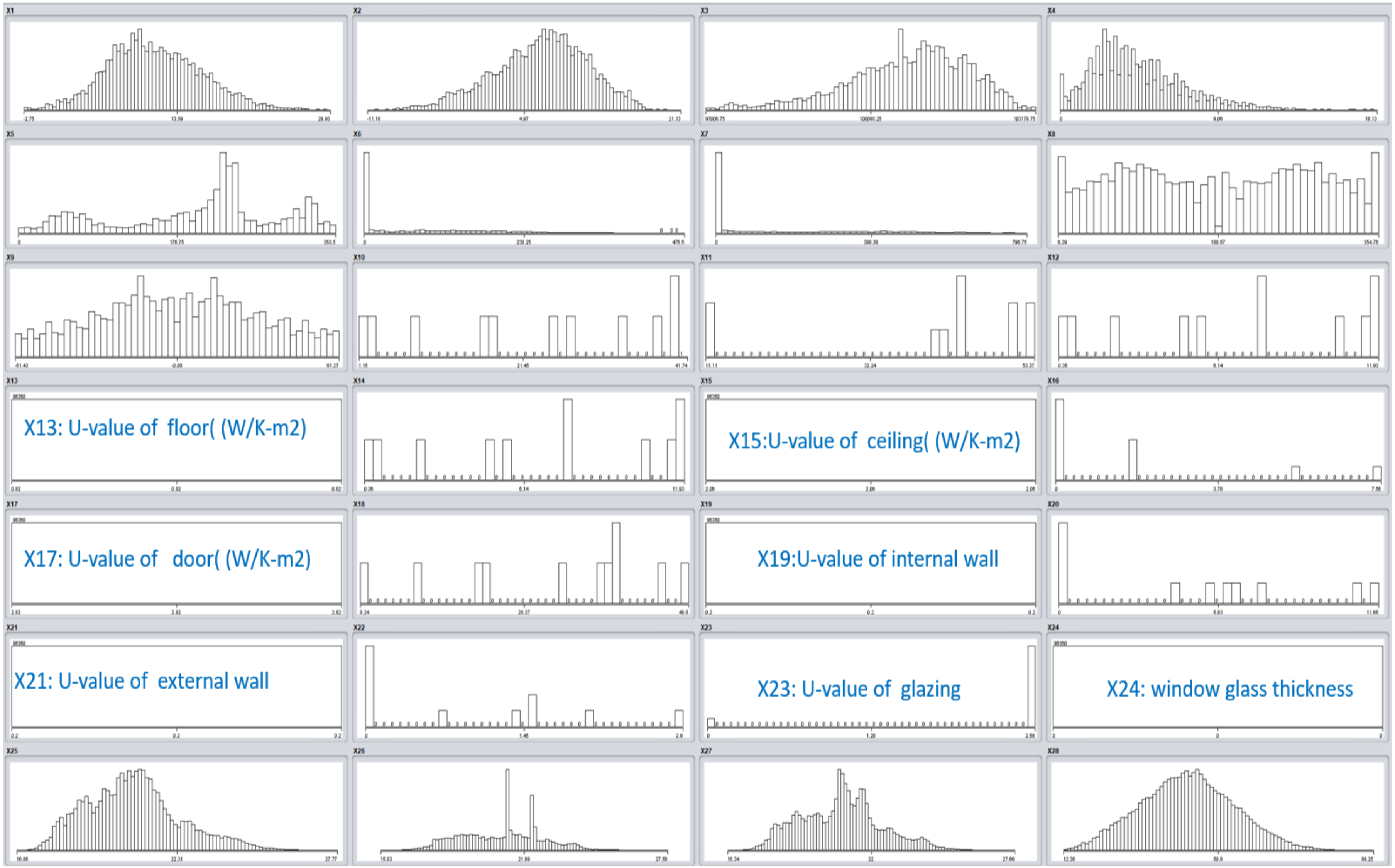


Figure 5. 16 Features distribution

5.7.2 Machine Learning Training

5.7.2.1 Linear Regression Model

The first machine learning algorithm employed in this study was a linear regression model. The goal was to predict thermal comfort, represented by the Predicted Mean Vote (PMV) index (Y1), using 29 input attributes (X1 to X28, plus Y1). A total of 96,360 instances were randomly selected from the dataset, which was then split into 70% for training and 30% for testing. The linear regression model produced the following equation:

$$Y1 = -0.0596 * X1 + 0.0712 * X2 -0 * X3 + 0.0099 * X4 + 0.0002 * X5 + 0.0014 * X6 + 0.0002 * X7 + 0.0013 * X8 -0.0125 * X9 -0.0336 * X12 -0.0336 * X14 + 0.2985 * X16 -0.0501 * X20 + 0.2621 * X22 + 0.5798 * X23 + 0.0654 * X25 -0.023 * X28 -2.8604$$

Key details of the linear regression training are as follows:

Time taken to build the model: 0.93 seconds

Time taken to test the model on the test split: 0.1 seconds

1.1.1.1 Evaluation Results of Linear Regression

The test dataset consisted of 28,908 instances, representing 30% of the original dataset. The evaluation metrics for the linear regression model are as follows:

- Correlation coefficient: 0.4696
- Mean absolute error (MAE): 0.7881
- Root mean squared error (RMSE): 0.9913
- Relative absolute error (RAE): 94.3412%
- Root relative squared error (RRSE): 88.2863%

The linear regression model yielded a correlation coefficient of 0.4696, indicating a moderate level of correlation between the input variables and the output (thermal comfort index: PMV). However, the high values for relative absolute error and root relative squared error suggest that the model may not have captured the complexity of the dataset effectively. This could be due to the limitations of linear regression in handling complex, non-linear relationships, which are likely present in the data.

5.7.2.2 Artificial Neural Network (ANN) - Multilayer Perceptron (MP)

For the second algorithm, an ANN-Multilayer Perceptron (MP) model was employed. The architecture consisted of 1 hidden layer with 14 neurons per layer (Figure 5.17). Key parameters for training the model were as follows:

- Number of epochs: 500

- Learning rate: 0.3
- Momentum: 0.2

The ANN-MP model is a more advanced algorithm capable of capturing non-linear relationships in the data, which is crucial for predicting variables like PMV that depend on complex environmental and structural factors. Key details of the ANN-MP training:

- Time taken to build the model: 287.54 seconds
- Time taken to test the model: 0.12 seconds

5.7.2.3 Evaluation Results of ANN-MP

The ANN model was also evaluated on the test dataset of 28,908 instances, and its performance was significantly better than that of linear regression. The results are as follows:

- Correlation coefficient: 0.9641
- Mean absolute error (MAE): 0.2242
- Root mean squared error (RMSE): 0.3057
- Relative absolute error (RAE): 26.8344%
- Root relative squared error (RRSE): 27.225%

The ANN-MP model demonstrated a strong correlation (0.9641) between the input features and the predicted PMV, showing that it effectively captured the complex relationships in the dataset. The lower MAE and RMSE values suggest that the ANN model provided more accurate predictions with fewer errors compared to the linear regression model.

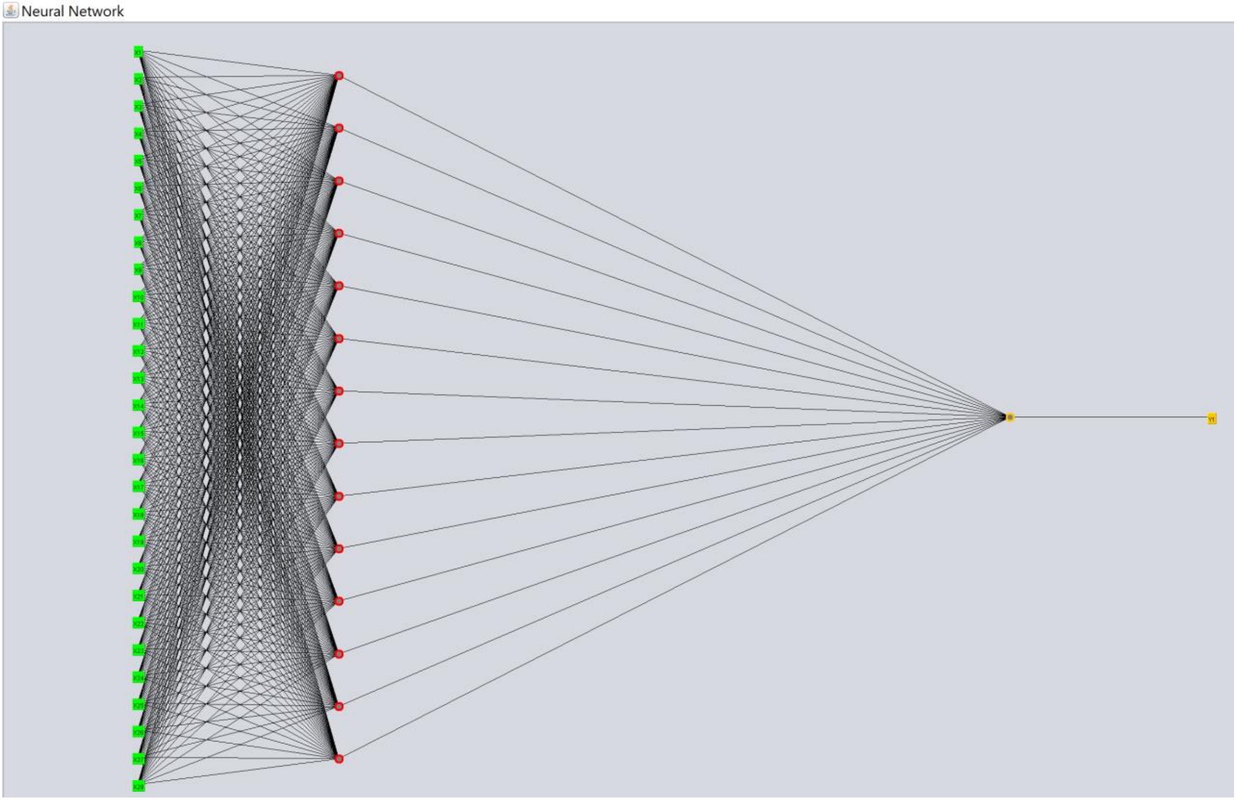


Figure 5. 17 The architecture of ANN-MP model (consisted of 1 hidden layer with 14 neurons per layer).

5.7.2.4 Random Forest - PMV Training

The Random Forest algorithm was applied next for PMV prediction. This ensemble learning technique utilized bagging with 100 iterations to improve the predictive accuracy.

Key details of the Random Forest training:

- Time taken to build the model: 40.08 seconds
- Time taken to test the model: 2.85 seconds

5.7.2.5 Evaluation Results of Random Forest (PMV)

The performance of the Random Forest model was outstanding, as demonstrated by the following metrics:

- Correlation coefficient: 0.9849
- Mean absolute error (MAE): 0.1087
- Root mean squared error (RMSE): 0.1977
- Relative absolute error (RAE): 13.0129%
- Root relative squared error (RRSE): 17.6082%

The Random Forest model achieved an almost perfect correlation coefficient of 0.9849, indicating that it was highly effective in capturing the relationships between input variables and PMV. The significantly lower error metrics (MAE, RMSE) compared to the ANN model suggest that Random Forest provided more accurate predictions with fewer errors.

5.7.2.6 Random Forest - PPD Training

Lastly, Random Forest was also used to predict the Predicted Percentage of Dissatisfied (PPD), another thermal comfort index. The same configuration was used with 100 iterations for bagging.

Key details of the Random Forest PPD training:

- Time taken to build the model: 41.7 seconds
- Time taken to test the model: 3.6 seconds

5.7.2.7 Evaluation Results of Random Forest (PPD)

The results for PPD prediction are:

- Correlation coefficient: 0.9799
- Mean absolute error (MAE): 3.6215
- Root mean squared error (RMSE): 6.7051
- Relative absolute error (RAE): 12.6195%
- Root relative squared error (RRSE): 20.4084%

The high correlation coefficient of 0.9799 demonstrates that Random Forest also performed well in predicting PPD. Although the MAE and RMSE were slightly higher compared to the PMV model, these values are still relatively low, confirming that Random Forest is a robust and reliable algorithm for predicting thermal comfort indices.

5.7.3 Summary of Findings

Random Forest - PMV Training: This model outperformed all others with the highest correlation coefficient (0.9849) and the lowest error rates across all metrics, making it the best algorithm for PMV prediction.

Artificial Neural Network (ANN): The ANN model performed well, but its errors (MAE, RMSE) were higher than those of the Random Forest, particularly in complex datasets like this one. While still a strong candidate, it requires more computational resources.

Random Forest - PPD Training: Random Forest also performed well for predicting PPD, with high accuracy and relatively low error rates, though the errors were larger than for PMV.

5.7.4 Conclusion

Random Forest is the superior algorithm for predicting both PMV and PPD thermal comfort indices. It achieved the highest accuracy, the lowest error rates, and was computationally efficient compared to the ANN and linear regression models.

5.8 Case Study 3 for Multi-Objective Training

In this section, the study focuses on identifying and analysing common dwelling types in the UK, which include detached houses, semi-detached houses, flats/apartments, terraced houses, townhouses, bungalows, and cottages. The purpose of this case study is to extend the research scope by incorporating a multi-objective training approach, targeting various energy consumption and comfort metrics simultaneously. By doing so, the study seeks to explore how different housing typologies can serve as effective models for machine learning (ML) training, specifically in terms of energy efficiency and occupant comfort.

5.8.1 Energy simulation of Building 0

The same model used in Case Study 1 is selected as the building 0 for Case Study 2. A Building Information Model (BIM) of a typical terraced house located in Cardiff was developed using DesignBuilder. This model represents a two-storey, single-family house, which is a common housing type in the UK. For consistency and comparative analysis, we continue to focus on the middle-terraced house in this row of buildings, which was also used in the previous case study.

Given that this case study involves multi-objective machine learning training, the simulation goes beyond daylight illuminance analysis. Additional energy metrics, such as heating and cooling loads, electricity usage, and solar radiation absorption, are also simulated and included in the model. These simulations provide a more comprehensive dataset that can serve as a robust foundation for the machine learning training model. The integration of various energy factors will allow the study to evaluate the overall energy performance of the building while considering the balance between occupant comfort and energy efficiency.

This Figure 5.18 represents the simulation output from EnergyPlus, detailing temperatures, heat gains, and energy consumption for a building over the course of a year, from January 1 to December 31. The figure is divided into multiple sections, each displaying different aspects of the building's energy performance and environmental conditions.

- Energy Consumption (Top Section)

The top section illustrates various components of energy consumption throughout the year. The most prominent line represents fuel consumption, which increases significantly during colder months, reflecting higher heating demand. Other energy components, such as system fans, system pumps, and auxiliary energy, show more consistent, lower levels of consumption. Peaks in these metrics coincide with seasonal changes, indicating varying energy needs based on external temperature fluctuations.

- Temperature Profiles (Second Section):

  This section presents several temperature-related variables, including indoor and outdoor air temperatures, radiant temperature, operative temperature, and outside dry-bulb temperature. The outside air temperature follows a typical seasonal pattern, with lower temperatures in winter and higher temperatures in summer. The indoor-related temperatures remain relatively stable, reflecting the operation of heating, ventilation, and air conditioning (HVAC) systems in maintaining a comfortable internal environment despite external variations.

- Heat Gains (Third Section):

  The third section captures the heat gains within the building, such as from external infiltration, external venting, general lighting, occupancy, and solar gains through windows. Solar gains, as indicated by the yellow line, increase significantly during the warmer months, leading to a corresponding reduction in heating needs. The sensible heating and cooling loads reflect the building's response to these external heat gains, adjusting accordingly to maintain thermal comfort.

- Total Heat Balance (Bottom Section):

  The bottom section illustrates the total heat loss from the building, including mechanical ventilation, natural ventilation, and infiltration. Higher heat losses are observed during the colder months, particularly in winter, where ventilation and infiltration contribute to the increased need for heating. The fluctuations in heat loss correspond to changes in outdoor conditions, with the heat loss peaking during periods of higher external air infiltration or mechanical ventilation requirements.

Overall, this simulation output provides a detailed view of how the building's energy systems respond to external environmental conditions, occupant behaviour, and heat gains throughout the year. By analysing these variables, a comprehensive understanding of the building’s energy performance can be achieved, which is crucial for optimizing energy efficiency and improving occupant comfort in future designs.

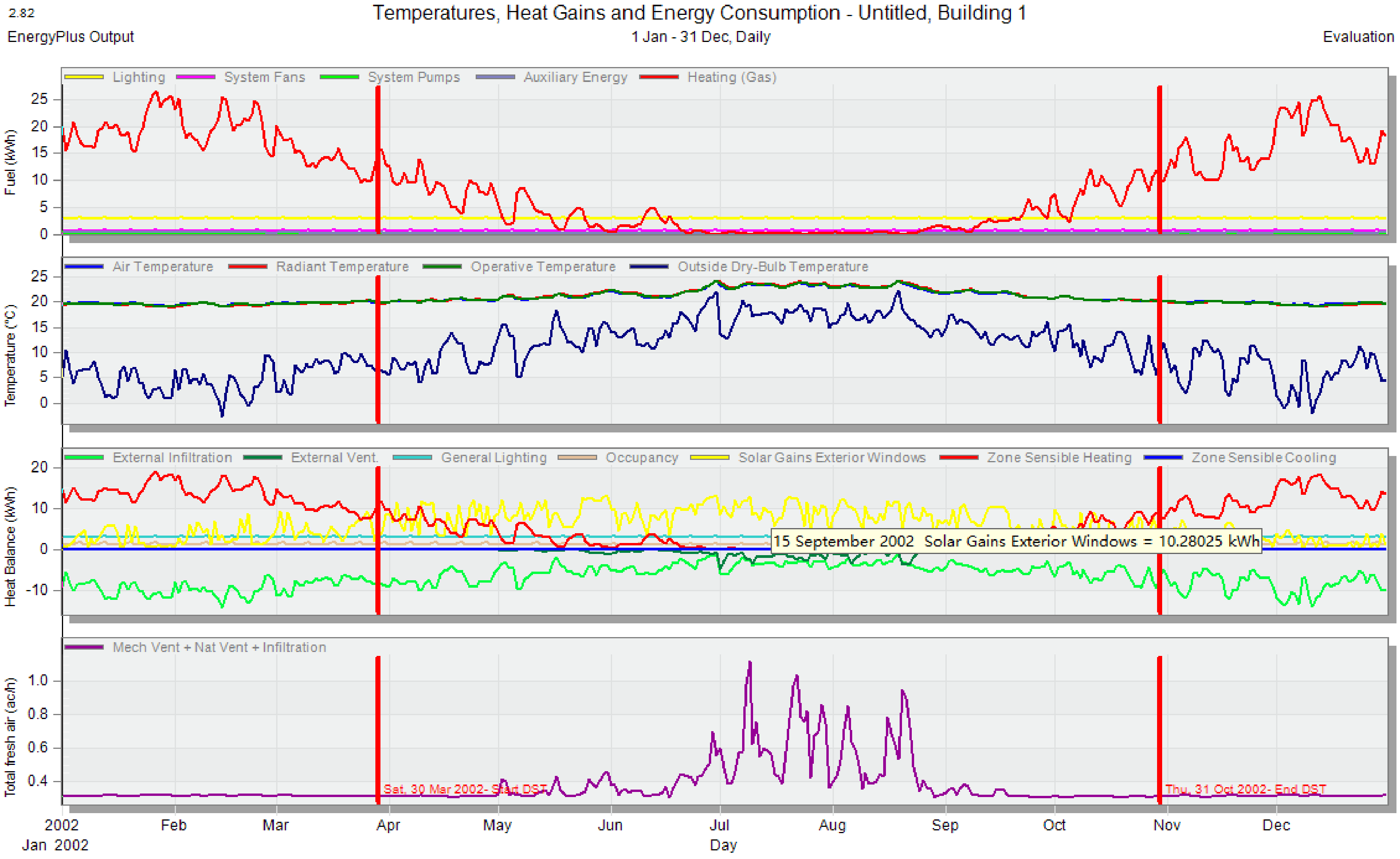


Figure 5. 18 Simulation outputs of building0 in DB

The Figure 5.19 depicts the hourly Predicted Percentage of Dissatisfied (PPD) values in the living room, based on the Fanger Thermal Comfort Model. PPD is a metric used to assess the percentage of people likely to feel uncomfortable in a given thermal environment. The horizontal axis represents the thermal comfort deviation (in unspecified units), while the vertical axis displays the PPD in percentage terms. As shown in the Figure 5.19, the PPD fluctuates along a curve, with the lowest values around thermal neutrality (closer to 0 on the horizontal axis), where dissatisfaction is minimal, representing optimal comfort. As the thermal conditions deviate from this neutral zone (either becoming too cold or too hot), the PPD increases significantly, indicating a higher percentage of people who are likely to be dissatisfied with the thermal conditions. When the deviation is at the extremes on both sides (around -4 and +4), the PPD reaches nearly 100%, reflecting almost total discomfort under those conditions. Conversely, in the central range, where the deviation is close to zero, the PPD decreases to values as low as 5-10%, indicating that the majority of people are likely to feel comfortable. This graph effectively illustrates the inverse relationship between the deviation from optimal thermal conditions and the predicted comfort of occupants in the living room, as modelled by the Fanger PPD approach.

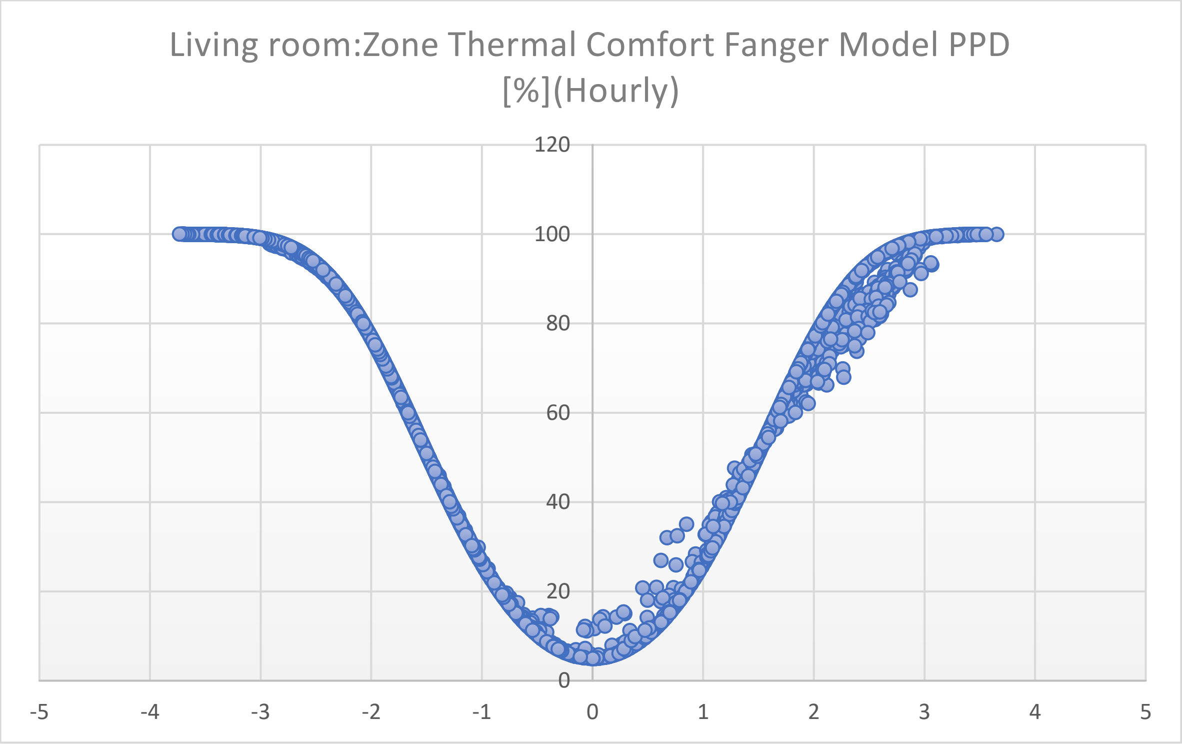


Figure 5. 19 Living room: Zone Thermal Comfort Fanger Model PPD [%](Hourly)

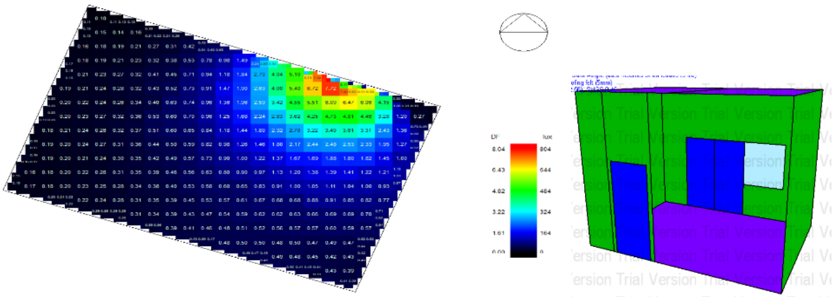


Figure 5. 20 1) Living room: daylight distribution in living room of building0; 2) living room model in design builder

5.8.2 Baseline Room Configuration

Building 0 contains hourly thermal comfort and illuminance data, with a total of 8,760 data points collected for a single room/unit over the course of a year. In total, there are 11 units in Building 0. If further simulations are conducted for other modified buildings, it becomes essential to systematically organize both the input and output data to facilitate efficient analysis. Given the typical dwelling types in the UK, the "box room" will be selected as the representative comfort unit for constructing a comprehensive database. This will serve as the baseline building model.

The overall building comfort is determined by aggregating the weighted ratios of various sub-room units within the building. Therefore, precise calculation of comfort levels for each individual unit is essential, as it serves as the foundation for assessing the comfort of the entire structure. In this study, the baseline building has been modelled in DesignBuilder, as illustrated in Figure 5.21. The room analysed in this study is a rectangular residential unit located in London, UK. It measures 3 meters in width, 6 meters in length, and 3.5 meters in height. The building is classified as a residential structure, with one exterior wall and the remaining walls functioning as internal partitions. This configuration serves as the basis for the comfort analysis and simulations conducted in the study. This setup provides the basis for conducting simulations and evaluating comfort metrics in a controlled, standardized environment, which can be used to assess various scenarios and modifications.

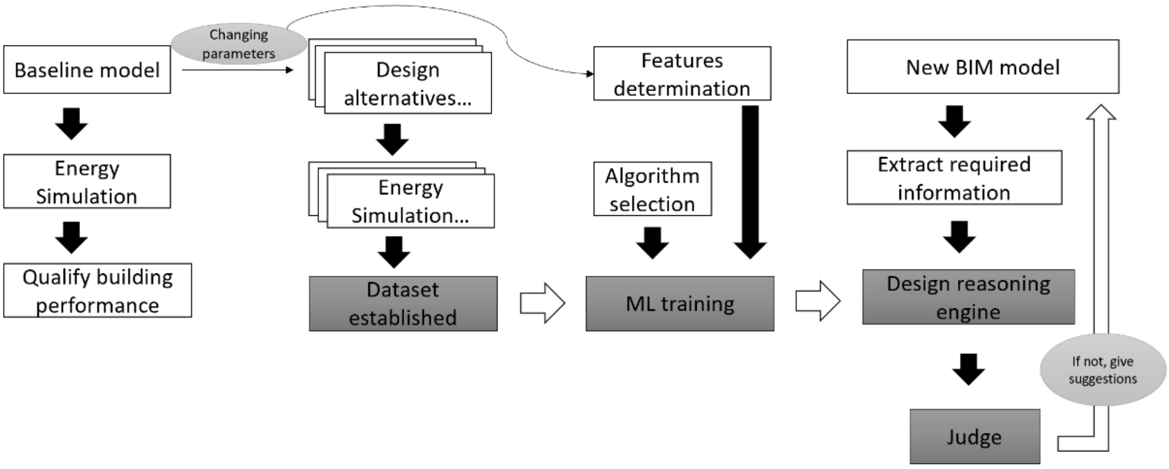


Figure 5. 21Workflow for ML in the context of energy simulation and performance optimization

The diagram above (Figure 5.21) illustrates the overall workflow for integrating Building Information Modelling (BIM) with machine learning (ML) in the context of energy simulation and performance optimization. Here are the several steps in this case study listed as follows.

- Baseline Model: The process begins with the creation of a baseline BIM model. This model represents the initial version of the building, incorporating key structural and design details. The baseline model is subjected to an energy simulation, which assesses the building's energy performance. This simulation provides essential insights into how the building performs under typical conditions. The results of the energy simulation are used to qualify the building's performance, including metrics related to energy efficiency, comfort, and other factors.
- Dataset Established: Following the baseline evaluation, various design alternatives are explored. These alternatives involve changing parameters within the building design (e.g., materials, window size, insulation) to generate new energy performance simulations. The energy simulation results from the different design alternatives are compiled to establish a comprehensive dataset. This dataset will be critical for training the ML model.
- ML Training: Features are determined from the dataset, and an appropriate ML algorithm is selected. The machine learning model is then trained on this dataset to learn patterns and make predictions about energy performance under various design configurations.
- Design Reasoning Engine: A new BIM model is created based on the suggestions from the ML model. The design reasoning engine, which is informed by the ML model's outputs, extracts the required information from the model for further evaluation.
- Judge and Feedback: The new BIM model is judged based on performance criteria. If the building performance meets the required standards, the design is finalized. If not, suggestions are provided, and the process is repeated to optimize the design.

This iterative approach helps optimize building design by leveraging machine learning and energy simulation to make data-driven design improvements.

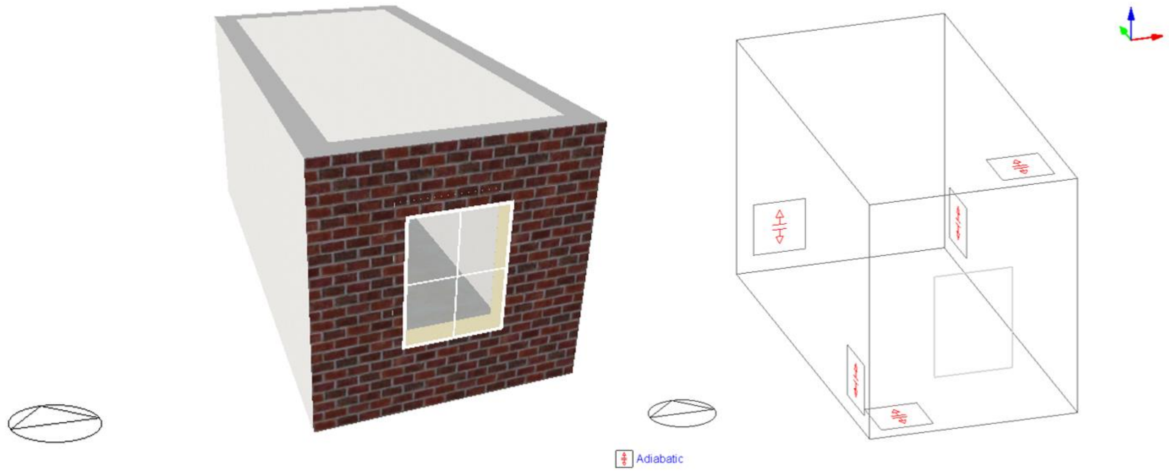


Figure 5. 22 The baseline building is built in DesignBuilder

5.8.3 Exploration of Design Alternatives

According to the overall workflow in this case study, Various design alternatives are explored in this section as shown in Figure 5.23. The dataset will be established by systematically altering the design alternatives.

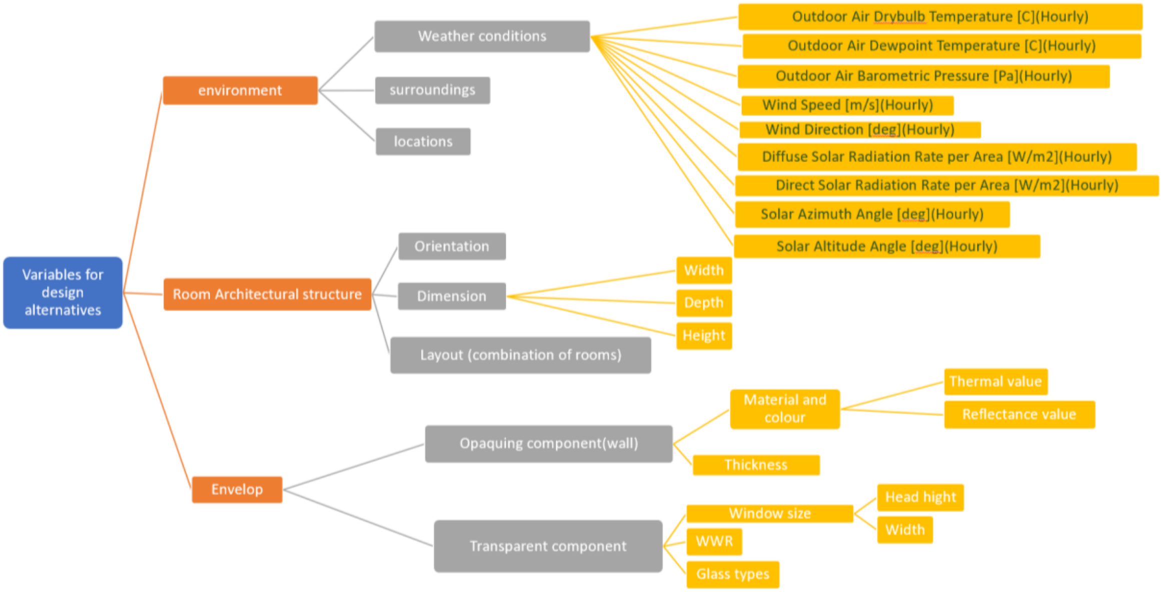

Variables for design alternatives
environment
Weather conditions
surroundings
locations
Outdoor Air Drybulb Temperature [C](Hourly)
Outdoor Air Dewpoint Temperature [C](Hourly)
Outdoor Air Barometric Pressure [Pa](Hourly)
Wind Speed [m/s](Hourly)
Wind Direction [deg](Hourly)
Diffuse Solar Radiation Rate per Area [W/m2](Hourly)
Direct Solar Radiation Rate per Area [W/m2](Hourly)
Solar Azimuth Angle [deg](Hourly)
Solar Altitude Angle [deg](Hourly)
Room Architectural structure
Orientation
Dimension
Layout (combination of rooms)
Width
Depth
Height
Envelop
Opaquing component(wall)
Material and colour
Thermal value
Reflectance value
Thickness
Transparent component
Window size
Head hight
Width
WWR
Glass types

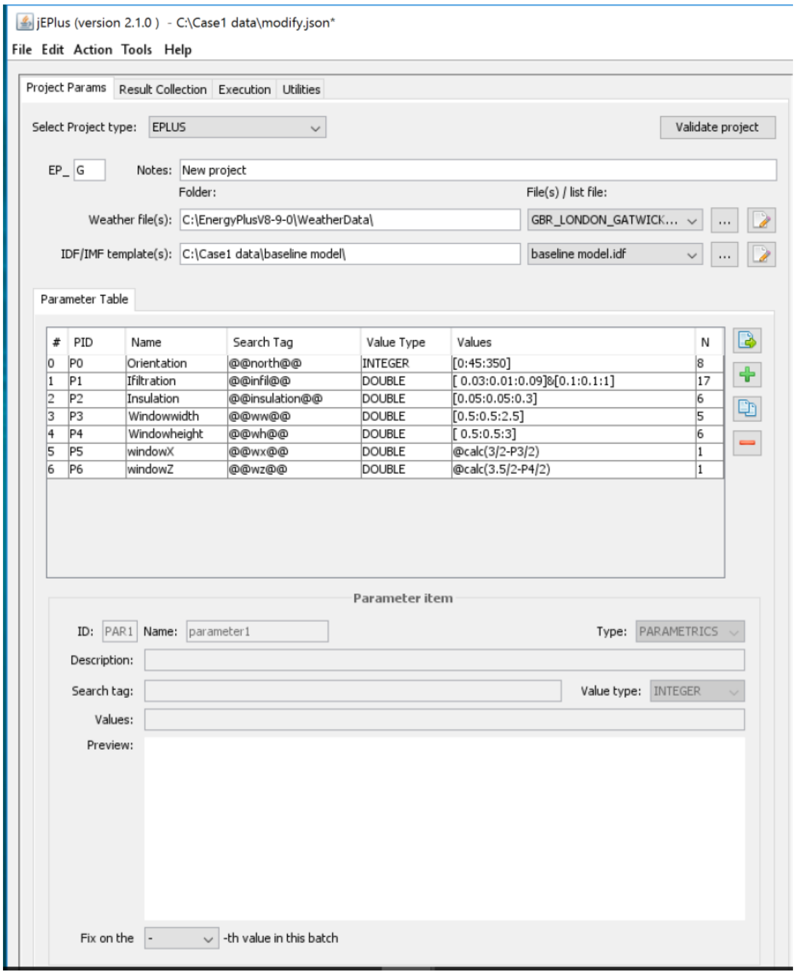
Figure 5. 23 Variables for design alternatives

The parametric simulations were managed using jEPlus, an open-source tool designed for running complex simulations with EnergyPlus and TRNSYS. The primary function of jEPlus is to facilitate the setup of parametric runs, allowing researchers and designers to explore how varying input parameters influence the overall building performance. By adjusting a broad set of design variables, jEPlus enables an in-depth analysis of factors such as energy consumption, thermal comfort, and daylighting, providing valuable insights for optimizing building design.

In this process, the input parameters could be modified, including but not limited to: building orientation, window-to-wall ratio, insulation levels, types of building materials. Each of these variables was iteratively adjusted and simulated using EnergyPlus through the jEPlus interface, yielding a large dataset for analysis and feeding into the machine learning model for further performance predictions.

Data is collected by changing the design alternatives. The primary purpose of jEPlus is to assist in setting up parametric runs with EnergyPlus models. This can be particularly useful for researchers or designers who are looking to explore a wide range of design variables to determine their impact on building performance. The input parameters can be changed from (Figure 5.23):

1. Orientation: {0, 45, 90,135,180,225,270,315}
2. Infiltration: {0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
3. Thickness of insulation layer: {0.05~0.3}
4. Window size and position modify:
    - Ww {0.5~2.5}
    - Wh {0.5~3}
    - (x,z)



Select Project type: EPLUS

EP_ G Notes: New project

Folder: File(s) / list file:

Weather file(s): C:\EnergyPlusV8-9-0\WeatherData\ GBR_LONDON_GATWICK...

IDF/IMF template(s): C:\Case1 data\baseline model\ baseline model.idf

#	PID	Name	Search Tag	Value Type	Values	N
0	P0	Orientation	@@north@@	INTEGER	[0:45:350]	8
1	P1	Ifiltration	@@infil@@	DOUBLE	[0.03:0.01:0.09]&[0.1:0.1:1]	17
2	P2	Insulation	@@insulation@@	DOUBLE	[0.05:0.05:0.3]	6
3	P3	Windowwidth	@@ww@@	DOUBLE	[0.5:0.5:2.5]	5
4	P4	Windowheight	@@wh@@	DOUBLE	[0.5:0.5:3]	6
5	P5	windowX	@@wx@@	DOUBLE	@calc(3/2-P3/2)	1
6	P6	windowZ	@@wz@@	DOUBLE	@calc(3.5/2-P4/2)	1

ID: PAR1 Name: parameter1 Type: PARAMETRICS

Description:

Search tag: Value type: INTEGER

Values:

Preview:

Fix on the - -th value in this batch

Figure 5. 24 The interface of jEplus

The dataset will be randomly selected for ML training. The figure 5.25 is one of the examples of modified room.

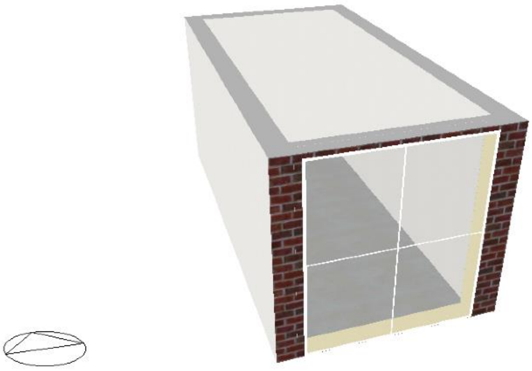


Figure 5. 25 The example of modified room

5.8.4 ML process

In this machine learning process, 1,000 instances were randomly generated using jEPlus to create a comprehensive dataset for analysis. The total time consumption for generating 1000 instances was more than 30 minutes, utilizing 2 threads for parallel processing on Virtual Machine in BIM lab of Cardiff University . The machine learning model was tested in a 66/34 split, with 66% of the dataset used for training and the remaining 34% for testing.

Input Parameters:

Since all alternative building designs were assumed to be located in London, weather-influencing variables were treated as constant, based on the annual average comfort levels per unit. The following seven input parameters were selected for this study:

- X_1 Orientation
- X_2 Infiltration
- X_3 Thickness of insulation layer
- X_4 Window width
- X_5 Window height
- X_6 Window left corner coordinate (X)
- X_7 Window left corner coordinate (Z)

Output Parameters:

The output parameters that were analysed include:

- Y_1 Discomfort hours (the number of hours in a year where the building is outside the comfort zone)
- Y2 Heating load and cooling load (annual energy consumption for heating and cooling the building)

5.8.5 Algorithm Selection and Results

5.8.5.1 Random Forest

The Random Forest algorithm was selected for its demonstrated ability to handle datasets with multiple parameters effectively. This method is particularly valued for its capability to avoid overfitting, its ensemble nature which ensures high accuracy, and its adaptability to various data types, making it a highly reliable choice for predictive modelling in this case. In this study, Random Forest was utilized to predict key performance metrics, including discomfort hours and heating/cooling loads. The algorithm's robustness was evident in its ability to maintain a balance between bias and variance, providing stable and precise predictions across different instances.

The key performance metrics for the Random Forest model were as follows:

- Correlation Coefficient (R^2): 0.9782

The high correlation coefficient reflects a strong positive relationship between the input parameters and the predicted output values, indicating that the Random Forest model was able to capture the underlying patterns in the data accurately.

- Mean Absolute Error (MAE): 114.3635
  MAE measures the average magnitude of errors in the predictions, without considering their direction. A lower value, such as in this case, demonstrates that the model made relatively small errors when predicting outcomes, which is crucial in energy simulations where minor discrepancies can significantly impact comfort levels.
- Root Mean Squared Error (RMSE): 147.657
  RMSE provides a measure of the average magnitude of the error, with a greater penalty for larger errors. The relatively low RMSE suggests that while some prediction errors exist, they are not substantial enough to reduce the model's reliability in predicting discomfort hours and energy loads.
- Relative Absolute Error (RAE): 21.6846%
  RAE normalizes the absolute error by the errors that would result from using a simple baseline model. The low percentage indicates that the Random Forest outperforms a naive approach by a significant margin.
- Root Relative Squared Error (RRSE): 22.2439%
  RRSE measures the efficiency of the model compared to a baseline model in predicting outcomes. The low RRSE indicates the model's high efficiency and accuracy relative to a baseline, further cementing Random Forest as a suitable model for this case study.

The overall performance of the Random Forest algorithm suggests that it is a highly effective method for predicting both discomfort hours and energy load in this specific building model. Its capacity to handle various parameters while providing a strong correlation with actual performance metrics makes it a strong candidate for future multi-objective optimization studies.

5.8.5.2 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) model was selected for its capability to model complex, non-linear relationships between input parameters and predicted outputs. Given the complexity of the dataset, ANN was chosen to explore its potential in capturing intricate patterns that simpler models, such as linear regression, may overlook. The primary objective was to determine the optimal configuration of hidden neurons in order to achieve the most accurate predictions for discomfort hours and energy loads.

The ANN model was tested with varying numbers of neurons in its hidden layer to identify the best performing architecture. Three configurations ANN were explored: 2 neurons; 4 neurons; 8 neurons. Each configuration involved a single hidden layer, while the input layer comprised the seven selected features (Orientation, Infiltration, Thickness of Insulation Layer, Window Width, Window Height, Window Left Corner X, Window Left Corner Z). The output layer provided predictions for discomfort hours and energy loads. The performance of each configuration was measured using several evaluation metrics, including the correlation coefficient, mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and root relative squared error (RRSE).

Results of ANN Configurations is showed in Figure 5.26. The results across different ANN configurations showed that the performance improved as the number of neurons increased. The key metrics for each configuration are as follows:

- 2 Neurons Configuration:
    - Correlation Coefficient: 0.908
    - Mean Absolute Error (MAE): 232.0217
    - Root Mean Squared Error (RMSE): 288.72
    - Relative Absolute Error (RAE): 43.9932%
    - Root Relative Squared Error (RRSE): 43.4944%

The configuration with 2 neurons provided reasonable accuracy but had higher error rates compared to the other configurations, making it less effective for this dataset.

- 4 Neurons Configuration:
    - Correlation Coefficient: 0.9774
    - Mean Absolute Error (MAE): 107.1521
    - Root Mean Squared Error (RMSE): 143.1205
    - Relative Absolute Error (RAE): 20.3169%
    - Root Relative Squared Error (RRSE): 21.5605%

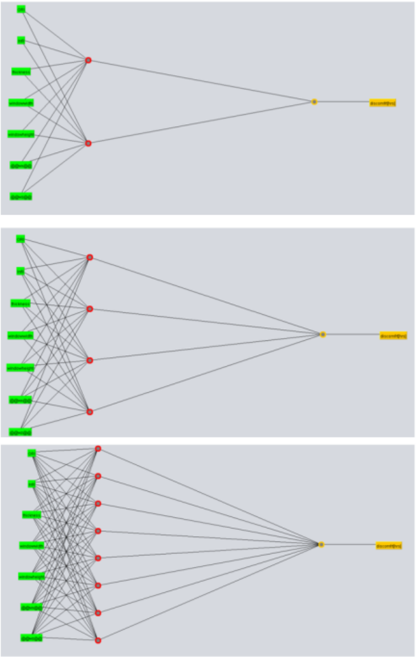
With 4 neurons in the hidden layer, the ANN model showed significant improvement. The correlation coefficient rose substantially, while MAE and RMSE were nearly halved, suggesting that this configuration captured the relationships between input and output variables more effectively.

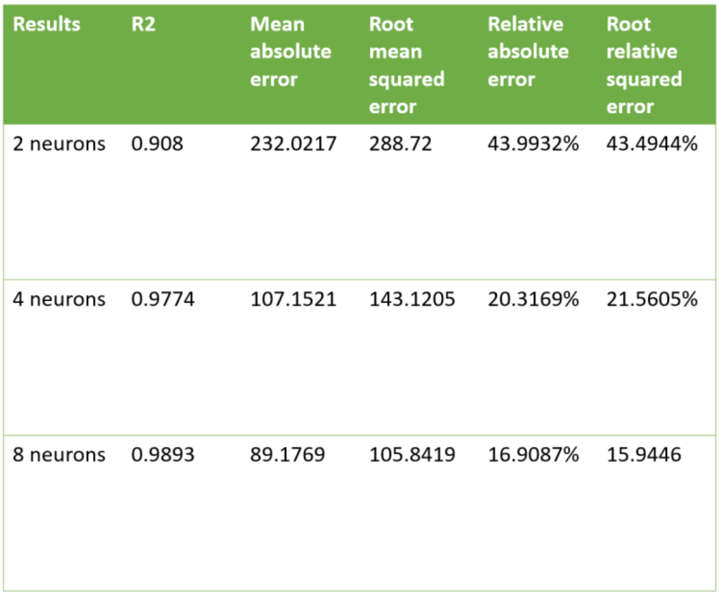
- 8 Neurons Configuration:
    - Correlation Coefficient: 0.9893
    - Mean Absolute Error (MAE): 89.1769
    - Root Mean Squared Error (RMSE): 105.8419
    - Relative Absolute Error (RAE): 16.9087%
    - Root Relative Squared Error (RRSE): 15.944%

The configuration with 8 neurons outperformed both the 2-neuron and 4-neuron configurations. It achieved the highest correlation coefficient (0.9893) and the lowest error metrics (MAE, RMSE, RAE, and RRSE), demonstrating its superior ability to model the complex relationships present in the dataset.

Among the three configurations tested, the 8-neuron configuration proved to be the most accurate, offering the highest correlation with the observed data and the lowest error metrics. The significant improvement in performance suggests that increasing the number of neurons in the hidden layer allows the ANN to model more intricate patterns, which is essential for predicting discomfort hours and energy loads in buildings.

The performance variation across different ANN configurations highlights the critical role of hyperparameter tuning—in this case, adjusting the number of neurons in the hidden layer—to optimize predictive performance. Testing different configurations ensures that the most suitable model is selected for the specific characteristics of the dataset, leading to more accurate and reliable predictions. The final model demonstrates that the Artificial Neural Network can effectively handle complex, non-linear relationships between input variables and predicted outcomes, making it a valuable tool in building energy consumption and comfort analysis.





Results	R2	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
2 neurons	0.908	232.0217	288.72	43.9932%	43.4944%
4 neurons	0.9774	107.1521	143.1205	20.3169%	21.5605%
8 neurons	0.9893	89.1769	105.8419	16.9087%	15.9446

Figure 5. 26 Results of ANN Configurations

5.8.6 Summary and Conclusion

Throughout this research, three distinct case studies were conducted, each focusing on different aspects of building performance and comfort prediction using machine learning (ML) models. These case studies employed various algorithms, such as Random Forest (RF), Artificial Neural Networks (ANN), and Linear Regression (LR), to evaluate and predict different building comfort indices, including daylight illumination, PMV-PPD, and discomfort hours.

The first case study focused on predicting daylight illuminance in a residential building. Using input variables such as outdoor air temperature, wind speed, solar radiation rates, and the position of windows, a Random Forest and Linear Regression model were developed to estimate indoor daylight levels. The Random Forest algorithm outperformed Linear Regression in both prediction accuracy and error minimization, achieving a higher correlation coefficient and lower mean absolute error (MAE) and root mean squared error (RMSE). The success of this case study demonstrated the efficacy of Random Forest in handling complex datasets, especially in predicting environmental factors like daylight, which are influenced by numerous variables.

The second case study aimed at predicting thermal comfort indices, specifically Predicted Mean Vote (PMV) and Predicted Percentage of Dissatisfied (PPD), using various machine learning algorithms. Both Random Forest and ANN were utilized, with Random Forest emerging as the superior model. It achieved higher accuracy, with a correlation coefficient of 0.9849 for PMV and 0.9799 for PPD, alongside lower error metrics compared to the ANN model. The performance of Random Forest highlighted its robustness in capturing the intricate relationships between variables affecting thermal comfort, such as air temperature, humidity, and building insulation. While ANN showed promise, it required significantly more time to train and exhibited slightly higher error rates.

The third case study shifted focus to predicting discomfort hours, which are the number of hours per year when indoor conditions fall outside of the comfort zone. In this case, Random Forest and ANN models were applied to a dataset with variables such as orientation, infiltration, insulation, and glazing. Again, the Random Forest model demonstrated superior performance, achieving better prediction accuracy and lower error metrics than the ANN model. This case study reinforced the advantage of using ensemble learning methods like Random Forest in handling large datasets with multiple variables that influence comfort over time.

Across all three case studies, Random Forest consistently outperformed other algorithms, including ANN and Linear Regression, in predicting building comfort metrics such as daylight illumination, PMV-PPD, and discomfort hours. The high accuracy and low error rates achieved by Random Forest across these diverse datasets highlight its flexibility and robustness in handling both environmental and thermal comfort variables.

While ANN also showed strong predictive capabilities, it required longer training times and was slightly less accurate in some cases, indicating that further tuning may be necessary to match the performance of Random Forest. Linear Regression, being a simpler model, struggled to capture the complexities of the datasets, leading to lower prediction accuracy and higher error rates.

In conclusion, Random Forest is the most reliable algorithm for predicting building comfort metrics in complex datasets, offering the best balance between prediction accuracy, computational efficiency, and error minimization. It proves to be an ideal choice for future applications in energy-efficient building design and real-time performance monitoring.

6 Development of a Customized Preference-Based Ontology for Comprehensive Indoor Comfort Assessment

The growing demand for personalized and adaptive living environments has underscored the necessity for advanced tools in the assessment and enhancement of indoor comfort. Traditional methods of comfort evaluation often fall short of addressing the complex interplay of factors that influence overall comfort. These conventional approaches typically rely on standardized indices, which fail to capture the nuances of individual preferences and the multi-dimensional nature of comfort. This research seeks to bridge this gap by developing a comprehensive ontology that not only evaluates multiple dimensions of comfort but also provides tailored improvement suggestions based on these assessments.

As smart building technologies advance, the concept of personalized indoor comfort has gained increasing importance. Conventional models, such as the Predicted Mean Vote (PMV) for thermal comfort, often overlook the variability in personal comfort thresholds and environmental interactions. Such models generally apply broad, one-size-fits-all criteria, which may not fully address individual needs or preferences. In response, this study introduces a user-oriented indoor comfort assessment framework based on ontology and the Semantic Web Rule Language (SWRL). By integrating multi-dimensional factors such as thermal comfort, acoustic comfort, visual comfort, and indoor air quality, this framework aims to provide a more customized and holistic approach to comfort evaluation, allowing for personalized design recommendations.

As introduced in Chapter 4, a comprehensive comfort framework was established, setting the foundation for this research. Traditional indoor comfort assessments often neglect the dynamic interplay of comfort dimensions and personal preferences, limiting their relevance in modern smart buildings. With the evolving understanding of comfort and the rise of smart technologies, there is an increasing need for adaptive, user-driven comfort assessments that can adjust based on occupant feedback and environmental data.

Furthermore, Chapter 5 highlighted the role of machine learning in facilitating the rapid calculation of comfort indices, significantly streamlining processes traditionally handled by energy simulation models. Machine learning has proven to be a valuable tool in providing faster, more accurate comfort assessments by replacing manual simulations with predictive algorithms.

The primary objective of this research is to develop an innovative ontology-based approach for indoor comfort assessment, designed to capture the complex relationships between various comfort factors and user preferences. Through this ontology, the method offers a structured and flexible framework capable of integrating multiple comfort dimensions, including real-time contextual data, thereby providing a more personalized and comprehensive evaluation of indoor comfort. By accounting for user-defined preferences, the system enhances the accuracy and relevance of the comfort assessment, offering tailored recommendations for design improvements.

This chapter outlines the design and development of the ontology, which integrates factors such as thermal, acoustic, visual, and air quality comfort. The proposed framework uses the Semantic Web Rule Language (SWRL) to enable dynamic, real-time assessments that adapt to evolving user needs and environmental conditions. Through this ontology, a new method for evaluating room comfort and suggesting design optimizations is introduced, providing a foundation for future advancements in personalized building comfort systems.

6.1 Introduction to Ontology in Comfort Assessment

The concept of ontology in building science refers to a structured framework that organizes information and relationships between different elements involved in indoor comfort. Ontologies can be used to define the relationships between parameters such as room temperature, humidity, airflow, daylight, noise levels, and the occupants' subjective preferences. By creating an ontology that integrates these factors, a comprehensive building comfort assessment system can be developed.

Key Components of the Ontology are displayed as follows according to the investigation in Chapter 4.

- Thermal Comfort: Parameters like temperature, humidity, airflow, and insulation affect thermal comfort.
- Visual Comfort: Daylight illuminance, glare, window design, and artificial lighting.
- Acoustic Comfort: Noise levels, sound insulation, and sources of noise.
- Air Quality: Indoor air quality, ventilation, CO2 levels, and pollutant concentrations.

6.2 Motivation for Developing a Customized Ontology

Traditionally, indoor comfort assessment has been carried out using standardized methods and generalized metrics, such as the Predicted Mean Vote (PMV) and Predicted Percentage Dissatisfied (PPD) models. While these models are useful for broader applications, they often overlook individual preferences and contextual differences between buildings and occupants. The limitations of generalized comfort models motivate the need for a more customizable approach.

Customization Based on User Preferences: Each occupant may have different comfort needs based on their activities, age, health status, and personal preferences. For example, one person may prioritize acoustic comfort in an office space, while another might place more emphasis on air quality in a residential setting. Developing an ontology based on these individual preferences enables a more personalized and accurate evaluation of indoor comfort.

6.3 Procedure of Implementation of Ontology

In this study, ontology is utilized to assess room comfort levels. Within information science and computer science, particularly in the domains of smart environments and intelligent buildings, employing ontology to represent and reason about the knowledge concerning environmental conditions, user preferences, and comfort standards proves to be an effective method. Here are the steps to implement room comfort determination using ontology (Lork et al., 2019):

1) Defining Comfort-Related Ontology: Initially, an ontology that includes all vital concepts and entities related to room comfort needs to be defined. This may encompass physical parameters such as temperature, humidity, light intensity, noise levels, and air quality, along with user preferences and types of activities.
2) Modelling Entities and Relationships: In the ontology, the relationships between these concepts must be explicitly described. For instance, user preferences might influence the comfort threshold definitions for specific environmental parameters. Similarly, the type of activity (such as working, resting, or reading) may dictate different comfort standards which showed in Figure 6.1 below. The proposed template for knowledge modelling is able to represent any building entities by different actors' perspective: architects as well as engineers up to client and users by using this model can represent entities in terms of 'their' meanings (specialist and maybe different for each of them), properties and rules, by this way being allowed to point out requirements and intents.

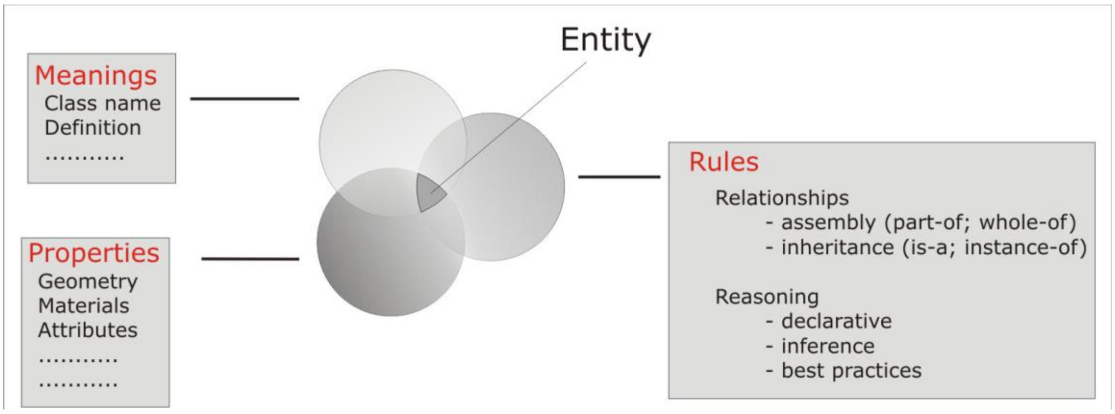


Figure 6. 1 The general template of knowledge representation (Carrara et al., 2009)

3) Data Collection and Integration: Abandoning traditional sensor-based real-time data collection of room environments, this study, building on the discussions in Chapter Four, employs simulated parameters. This approach effectively allows for preliminary assessments of building comfort during the early stages of architectural design and integrates this data with the ontology.
4) Reasoning and Decision Making: Utilizing an ontology-based reasoning engine enables

complex logical determinations to ascertain whether the current design meets the user's comfort requirements. If the design falls short of the standards, the system can automatically suggest adjustments to architectural parameters or recommend that users modify environmental settings during the operational phase of the building, such as adjusting temperature, humidity, or lighting to enhance comfort.

5) User Feedback and Adaptive Learning: By collecting user feedback, the user preference model within the ontology can be further refined, enabling the system to learn and adapt to the comfort needs of users.

6.4 CBContology Design

The customized ontology for indoor comfort assessment is structured into four primary categories: thermal comfort, visual comfort, acoustic comfort, and indoor air quality. Each category encompasses multiple sub-categories that represent the specific factors influencing comfort within each domain. This structured ontology provides the foundation for a comprehensive assessment and recommendation process tailored to the unique comfort needs of building occupants.

6.4.1 Ontology Structure:

Thermal Comfort:

- Room temperature (set point, variations)
- Humidity levels
- Airflow (distribution through HVAC systems)
- Building materials (thermal insulation properties)

Visual Comfort:

- Daylight availability (windows, shading devices)
- Artificial lighting (lamp type, intensity)
- Glare (from external and internal sources)

Acoustic Comfort:

- Background noise levels
- Sound insulation (partition materials, glazing)
- Sources of noise (internal and external)

Air Quality:

- Ventilation rates
- Indoor pollutant levels (CO2, VOCs)

- Filtration systems (HVAC filters)

6.4.2 System Framework and Vital Components

The Comprehensive Building Comfort Ontology (CBContology) is designed with four key components that work together to evaluate and enhance indoor comfort effectively:

- Knowledge Base:
  The knowledge base serves as the central repository, storing all ontology models and rules developed using the Semantic Web Rule Language (SWRL). It contains the core information, including all comfort dimensions, associated parameters, and their interrelationships. This component is essential for ensuring that all relevant comfort factors are captured and structured for the system's operational processes.
- Ontology Management System:
  The ontology management system enables the creation, modification, and management of the ontology. Protégé, a widely used open-source tool, was selected for this study. Protégé offers a flexible platform for building complex ontologies and managing the relationships between different knowledge elements. Its compatibility with SWRL and its user-friendly interface make it an ideal choice, particularly given the need to model the intricate relationships among comfort factors such as thermal, acoustic, visual, and air quality dimensions.
- Rule Engine:
  The rule engine is responsible for interpreting the facts and SWRL rules stored within the knowledge base. By processing the ontology and predefined rules, it generates new facts and insights. This dynamic reasoning capability allows the system to assess real-time data and offer targeted recommendations to improve indoor comfort, making it a crucial component for adaptive decision-making.
- Query Interface:
  The query interface facilitates interaction between users and the ontology system. Users can input preferences or comfort-related queries, and the system responds by retrieving relevant data and providing customized insights or suggestions. This component ensures that the ontology remains responsive to the specific needs of occupants or building designers, offering a user-centred approach to comfort assessment.

These four components work in conjunction, ensuring the ontology system functions seamlessly and efficiently. Additionally, a reasoning engine is embedded in the system to check the consistency of the developed ontology. This consistency check helps identify and rectify potential errors, ensuring the ontology remains valid and capable of supporting accurate comfort assessments.

6.4.3 The Reason for Choosing of Protégé

Protégé was selected for this project due to its versatility and robustness in supporting complex ontology development. As an open-source tool, it provides a reliable platform for both academic research and practical applications. Protégé's compatibility with SWRL allows for the seamless integration of rules necessary for reasoning about the various dimensions of comfort. Additionally, its active developer community ensures continuous updates, making it a dependable choice for managing the evolving demands of smart building technologies.

The tool's intuitive user interface makes it easy to build, manage, and extend ontologies, which is particularly important given the multi-dimensional nature of indoor comfort. Protégé simplifies the handling of interrelated comfort factors such as thermal conditions, air quality, visual comfort, and acoustic elements, ensuring that the developed ontology can accommodate a wide range of user preferences and environmental variables.

6.5 Methodology

To establish a knowledge-based decision support system, it is first necessary to identify the complex field knowledge before incorporating it into the knowledge base. Collecting domain knowledge is a vital preparation step in the creation of an ontology.

To acquire domain knowledge, a large number of knowledge engineering methods have been developed, for example, MIKE, CommonKADS, and PROTEGE-II. However, each of them has employed a different emphasis (Hou et al. 2015).

CommonKADS by Schreiber (2000) primarily consists of three activities:

- Knowledge identification. It involves identifying the problems in relevant domains, the purpose the knowledge will serve, and the scope of the ontology.
- Knowledge specification. During this step, a template is selected, and a semi- formal model is developed. The purpose of these activities is to create a specification for the knowledge model.
- Knowledge refinement. It is the last phase of the knowledge modelling process, and it typically consists of two tasks: validation and refinement of the knowledge model.

6.5.1 Development of CBContology

Based on the acquired domain knowledge, knowledge engineers can start to develop the ontology. According to the previous literature, several approaches can be adopted to develop an ontology, such as Uschold and King, Grüninger and Fox, Methodology, KACTUS and Ontology Development 101(Noy et al. 2001).

Ontology Development 101 is adopted due to the following reasons:

  - (1) This methodology was designed for beginners. As such, it is easy to learn and operate.
  - (2) The detailed activities involved in this approach have been specified. The process of establishing an ontology is described in detail in this methodology.
  - (3) It can be integrated with other tools. This method contains detailed instructions on how to implement the ontology in the Protégé environment.

Here is the key step of development of CBContology

Step 1 Determine the domain and scope of the CBContology

Basic questions and answers:

- Why develop the CBContology?
  - To improve the management of multi-domain knowledge, to help architects reduce repetitive work tasks, and provide more valuable information to support effective decision making in design stage.
- What is the domain that the ontology will cover?
  - Building performance (comfort condition and energy efficient) and building design (building architecture model)
  - Comfort: thermal sensation (to reach thermal neutrality) and maximise the daylight illuminance value.
- For what we are going to use the ontology?
  - Make judgment for comfort conditions according to the room information from BIM model in design stage. Using the ontology technology to build a knowledge-based suggestion engine which can replace the experts from energy and architecture domain to share and communication For enhance the information exchange between energy experts and architect to realise the automatically process engine. This suggestion engine can provide the corresponding design suggestions according to the results from ML process.
- For what types of questions, the information in the ontology should provide answers?
  - Why the room is discomfort? (General reasons)
  - What design suggestion can be given for this room? (according to different comfort aspect, improve the design)
- Who will use and maintain the ontology?
  - The expert in building performance domain and architect agency.

Step 2. Consider reusing existing ontologies

There are libraries of reusable ontologies on the Web and in the literature. For example, the Ontolingua ontology library (http://www.ksl.stanford.edu/software/ontolingua/) or the DAML ontology library (http://www.daml.org/ontologies/). There are also a number of publicly available commercial ontologies. However, no relevant ontologies already exist and start developing the ontology from scratch.

Step 3. Enumerate important terms in the ontology to write down a list of all terms we would like either to make statements about or to explain to a user.

Here is a Figure 6.2 show essential terms related to room comfort evaluation and building design.

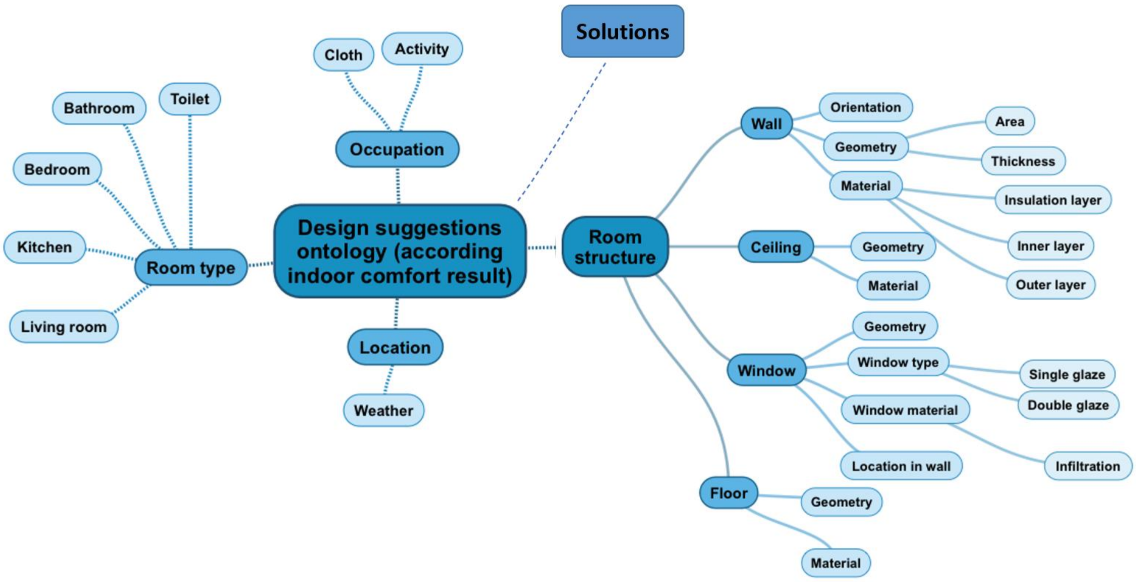


Figure 6. 2 Detailed classes and class hierarchy in CBContology

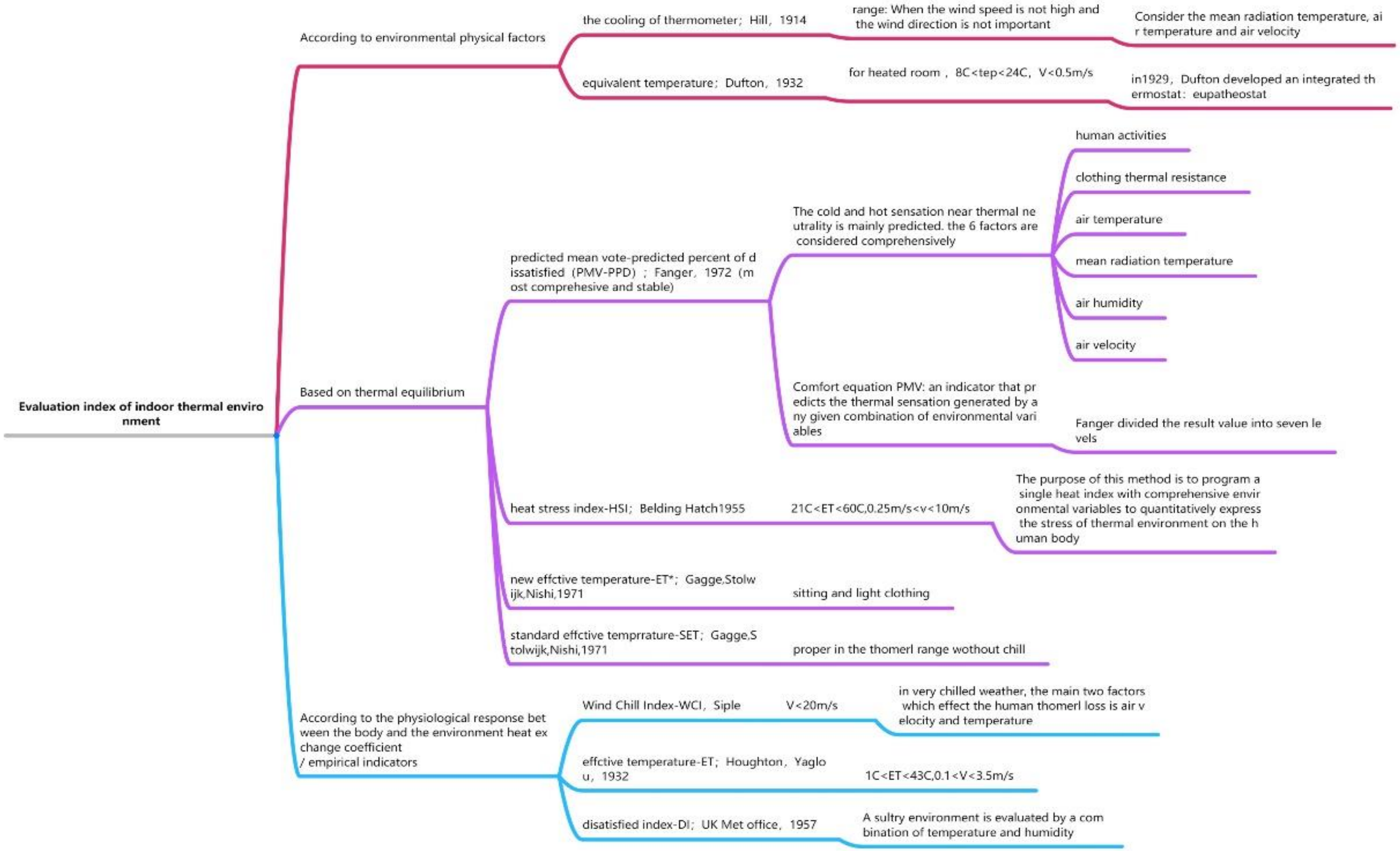

Evaluation index of indoor thermal enviro nment
According to environmental physical factors
the cooling of thermometer；Hill，1914
range: When the wind speed is not high and the wind direction is not important
Consider the mean radiation temperature, ai r temperature and air velocity
equivalent temperature；Dufton，1932
for heated room，8C<tep<24C，V<0.5m/s
in1929，Dufton developed an integrated th ermostat：eupatheostat
Based on thermal equilibrium
predicted mean vote-predicted percent of d issatisfied（PMV-PPD）；Fanger，1972（m ost comprehesive and stable)
The cold and hot sensation near thermal ne utrality is mainly predicted. the 6 factors are considered comprehensively
human activities
clothing thermal resistance
air temperature
mean radiation temperature
air humidity
air velocity
Comfort equation PMV: an indicator that pr edicts the thermal sensation generated by a ny given combination of environmental vari ables
Fanger divided the result value into seven le vels
heat stress index-HSI；Belding Hatch1955
21C<ET<60C,0.25m/s<v<10m/s
The purpose of this method is to program a single heat index with comprehensive envir onmental variables to quantitatively express the stress of thermal environment on the h uman body
new effctive temperature-ET*；Gagge,Stolw ijk,Nishi,1971
sitting and light clothing
standard effctive temprrature-SET；Gagge,S tolwijk,Nishi,1971
proper in the thomerl range wothout chill
According to the physiological response bet ween the body and the environment heat ex change coefficient / empirical indicators
Wind Chill Index-WCI，Siple
V<20m/s
in very chilled weather, the main two factors which effect the human thomerl loss is air v elocity and temperature
effctive temperature-ET；Houghton，Yaglo u，1932
1C<ET<43C,0.1<V<3.5m/s
disatisfied index-DI；UK Met office，1957
A sultry environment is evaluated by a com bination of temperature and humidity

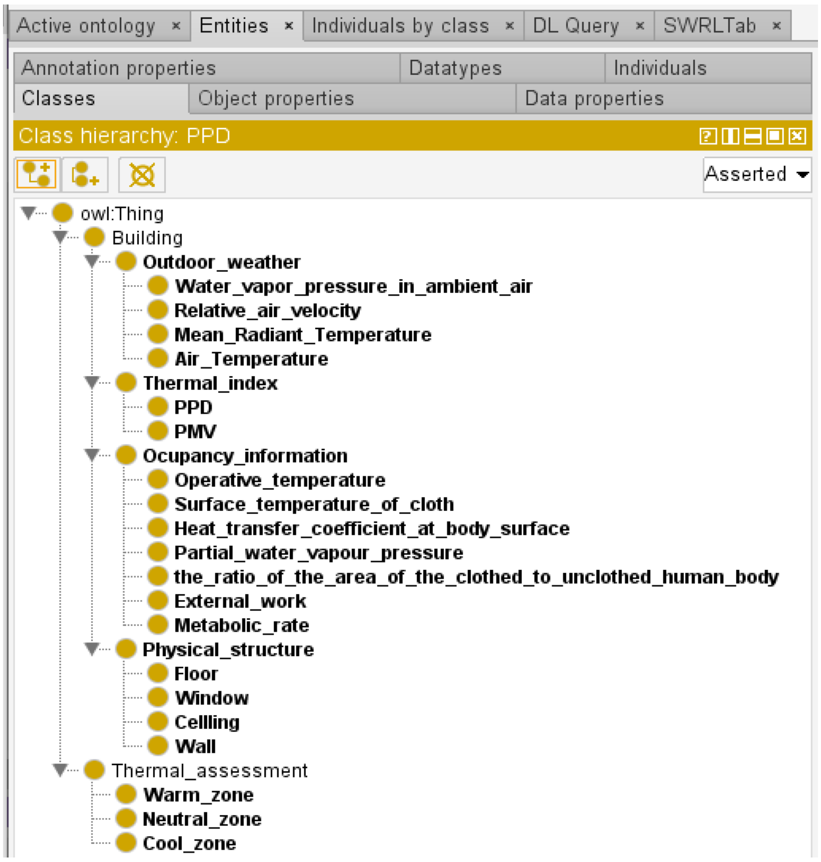
Figure 6. 3 Essential terms related to room comfort evaluation and building design

Step 4. Define the classes and the class hierarchy There are several possible approaches in developing a class hierarchy (Uschold and Gruninger 1996):

- A top-down development process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts.
- A bottom-up development process starts with the definition of the most specific classes, the leaves of the hierarchy, with subsequent grouping of these classes into more general concepts
- A combination development process is a combination of the top-down and bottom up approaches: The more salient concepts are defined first, followed by appropriate generalization and specialization.

Step 5. Define the properties of classes—slots

A class hierarchy in isolation cannot represent knowledge accurately; the properties of classes are also incorporated. There are three different kinds of properties are used in ontologies: object properties, data properties, and annotation properties. The Figure 6.4 below shows structured CBContology for building comfort assessment in protégé. This structure is part of a system that leverages these relationships for automated, ontology-based comfort evaluations.



Class hierarchy: PPD

- owl:Thing
  - Building
    - Outdoor_weather
      - Water_vapor_pressure_in_ambient_air
      - Relative_air_velocity
      - Mean_Radiant_Temperature
      - Air_Temperature
    - Thermal_index
      - PPD
      - PMV
    - Ocupancy_information
      - Operative_temperature
      - Surface_temperature_of_cloth
      - Heat_transfer_coefficient_at_body_surface
      - Partial_water_vapour_pressure
      - the_ratio_of_the_area_of_the_clothed_to_unclothed_human_body
      - External_work
      - Metabolic_rate
    - Physical_structure
      - Floor
      - Window
      - Cellling
      - Wall
  - Thermal_assessment
    - Warm_zone
    - Neutral_zone
    - Cool_zone

Figure 6. 4 Interface of protégé

Step 6. Define the facets of the slots

Slots can have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take.

- Slot cardinality：Slot cardinality defines how many values a slot can have
- Slot-value type： A value-type facet describes what types of values can fill in the slot.
- Common value types: String, Number, Boolean, Enumerated, Instance
- Domain and range of a slot： Allowed classes for slots of type Instance are often called a range of a slot.

Step 7. Create instances

The last step is creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires choosing a class, creating an individual instance of that class, and filling in the slot values.

6.6 SWRL Rule Formulation

In Chapter 4, the methodology for calculating the Comprehensive Building Comfort (CBC) value was outlined. Building upon that, in this chapter, the CBContology is designed to integrate the key comfort factors and enable dynamic comfort assessment. The CBC ontology is built on a comprehensive framework that accounts for Thermal Comfort (T), Acoustic Comfort (A), Indoor Air Quality (Q), and Visual Comfort (V), each of which is associated with specific attributes within the ontology to reflect their influence on overall room comfort.

6.6.1 Classes and Properties

Classes: Defined as ComfortAssessment, which encapsulates all the relevant data for assessing comfort in a room or building.

Properties: Include measurable attributes such as hasThermalComfortValue, hasAcousticComfortValue, hasIAQValue, and hasVisualComfortValue, corresponding to the key comfort dimensions. Additional properties, such as hasAlpha1 to hasAlpha4, represent the weighting coefficients for each comfort factor, enabling personalized or standardized weighting schemes.

6.6.2 SWRL Rule Formulation

To dynamically assess room comfort based on the defined factors, SWRL has been employed to articulate the calculation:

$$CBC = \alpha_1 T + \alpha_2 A + \alpha_3 Q + \alpha_4 V$$

This rule calculates the Comprehensive Building Comfort (CBC) value by integrating individual comfort parameters and weighting them based on their significance. The SWRL rule is expressed as follows:

```
ComfortAssessment(?ca) ^
hasThermalComfortValue(?ca, ?ther) ^
hasAcousticComfortValue(?ca, ?acou) ^
hasIAQValue(?ca, ?iaq) ^
hasVisualComfortValue(?ca, ?vis) ^
hasAlpha1(?ca, ?alpha1) ^
hasAlpha2(?ca, ?alpha2) ^
hasAlpha3(?ca, ?alpha3) ^
hasAlpha4(?ca, ?alpha4) ^
swrlb:multiply(?therMul, ?ther, ?alpha1) ^
swrlb:multiply(?acouMul, ?acou, ?alpha2) ^
swrlb:multiply(?iaqMul, ?iaq, ?alpha3) ^
swrlb:multiply(?visMul, ?vis, ?alpha4) ^
swrlb:add(?tempSum, ?therMul, ?acouMul) ^
swrlb:add(?tempSum2, ?tempSum, ?iaqMul) ^
swrlb:add(?cbc, ?tempSum2, ?visMul)
-> hasCBCValue(?ca, ?cbc)
```

ComfortAssessment(?ca) refers to the entity for which the comfort is being assessed. ***hasThermalComfortValue, hasAcousticComfortValue, hasIAQValue, hasVisualComfortValue*** represent the respective comfort values.

hasAlpha1, hasAlpha2, hasAlpha3, hasAlpha4 correspond to the weightings of each comfort factor.

The SWRL operations (multiply and add) compute the weighted sum of the comfort factors, resulting in the final CBC value.

6.7 Customised Comfort Preference

To develop a customized comfort assessment framework based on user preferences, an Analytical Hierarchy Process (AHP) was employed to determine the relative weights of thermal, acoustic, visual, and air quality comfort factors, which already presented in chapter 4. This approach allows users to assign importance levels to these four comfort dimensions based on personal preferences. The methodology involves the following steps:

- Determine Client Preferences: Users specify their preferences for the four comfort factors by providing a relative importance value, u_{ij}, between different factors, as per the AHP method.
- Construct the Evaluation Matrix (P): A 4x4 matrix P is created, where each element u_{ij} represents the importance of comfort factor u_i relative to factor u_j, based on user input.
- Calculate Eigenvector (ξ): The eigenvector corresponding to the largest eigenvalue of the evaluation matrix P is computed, yielding the relative weights of each comfort factor.
- Construct the Integrated Comfort Framework (C): The weighted comfort framework C is constructed using the computed weight vector and the time fractions of comfort for each factor (thermal, visual, acoustic, and air quality).

$$C = w_1 T + w_2 V + w_3 N + w_4 Q$$

Where T, V, N, and Q represent the time fractions during which comfort is achieved in a given period for thermal, visual, acoustic, and air quality dimensions, respectively.

Performing Analytic Hierarchy Process (AHP) operations typically involves dealing with matrix operations, including steps such as constructing judgement matrices, calculating weights, and consistency tests. While it is possible to perform these calculations manually, efficiency and accuracy can be greatly improved by using a programming language. Python is ideal for performing these types of mathematical and statistical operations because it has powerful mathematical and scientific computation libraries such as NumPy and SciPy. Below is a simple example Python script that demonstrates how to perform an AHP calculation using the NumPy library:

```
import numpy as np
def ahp_calculation(judgement_matrix):
    # Calculate eigenvalues and eigenvectors
    eigenvalues, eigenvectors = np.linalg.eig(judgement_matrix)
    # Find the eigenvector corresponding to the largest eigenvalue
    max_index = np.argmax(eigenvalues)
    max_eigenvector = eigenvectors[:, max_index].real
    # Normalize eigenvectors as weights
    weights = max_eigenvector / np.sum(max_eigenvector)
    # Consistency test (optional)
    CI = (eigenvalues[max_index] - len(judgement_matrix)) / (len(judgement_matrix) - 1)
    RI = [0, 0, 0.58, 0.9, 1.12, 1.24, 1.32, 1.41, 1.45]
    # Random Consistency Index based on matrix size
    CR = CI / RI[len(judgement_matrix)-1]
    print(f"Consistency Ratio (CR): {CR}")
    if CR < 0.1:
        print("Consistency accepted.")
    else:
        print("Consistency unacceptable, please re-evaluate the judgement matrix.")
    return weights
```

This Python script facilitates the AHP operations by calculating the eigenvalues, eigenvectors, and weights for the judgement matrix and performs a consistency check to ensure that the matrix has acceptable consistency.

np.linalg.eig is used to compute the eigenvalues and eigenvectors of the matrix.

The eigenvectors correspond to the maximum eigenvalues are normalised and output as the weights of the AHP. The consistency test is done by calculating the Consistency Index (CI) and Consistency Ratio (CR), which helps to assess the degree of consistency of the judgement matrix. If the CR is less than 0.1, the judgement matrix is considered to have satisfactory consistency. This script is a basic example of an AHP calculation; in practice, the construction of judgement matrices needs to be done based on user input or expert evaluation.

6.7.1 SWRL Rule Design

The SWRL rules can account for two cases: one where all the comfort factor weights are equal (0.25), and another where the weights are calculated dynamically based on user preferences using the AHP method.

6.7.1.1 Case 1: No User Preference (Equal Weights)

In this case, all comfort factors are assigned equal weight:

```
ComfortAssessment(?ca) ->
hasThermalWeight(?ca, 0.25) ^
hasAcousticWeight(?ca, 0.25) ^
hasVisualWeight(?ca, 0.25) ^
hasAirQualityWeight(?ca, 0.25)
```

6.7.1.2 Case 2: User Preferences (AHP-Calculated Weights)

When user preferences are provided, and weights are calculated using the AHP method, the rule for integrated comfort assessment is as follows:

```
ComfortAssessment(?ca) ^
hasThermalComfortScore(?ca, ?t) ^
hasAcousticComfortScore(?ca, ?a) ^
hasVisualComfortScore(?ca, ?v) ^
hasAirQualityScore(?ca, ?q) ^
hasThermalWeight(?ca, ?wt) ^
hasAcousticWeight(?ca, ?wa) ^
hasVisualWeight(?ca, ?wv) ^
hasAirQualityWeight(?ca, ?wq) ^
swrlb:multiply(?tempScore, ?t, ?wt) ^
swrlb:multiply(?acouScore, ?a, ?wa) ^
swrlb:multiply(?visScore, ?v, ?wv) ^
swrlb:multiply(?airScore, ?q, ?wq) ^
swrlb:add(?partialSum1, ?tempScore, ?acouScore) ^
swrlb:add(?partialSum2, ?partialSum1, ?visScore) ^
swrlb:add(?totalComfortScore, ?partialSum2, ?airScore)
-> hasTotalComfortScore(?ca, ?totalComfortScore)
```

This rule dynamically calculates the weighted sum of the comfort scores based on the user-provided weights, providing a flexible framework for assessing room comfort tailored to individual preferences.

By combining ontology and SWRL rules with the AHP method, this framework allows for dynamic and customizable comfort assessment. Whether considering equal weighting or user-defined preferences, this approach offers a flexible and powerful mechanism to assess and improve indoor comfort across various building types, enhancing both living and working environments.

6.8 Integrating Machine Learning for Real-Time Comfort Adjustments

Once the ontology is developed, integrating it with machine learning (ML) algorithms allows for dynamic, real-time optimization of indoor comfort conditions. This process leverages real-time environmental data and occupant feedback to adjust comfort settings continuously, ensuring that the indoor environment remains aligned with user preferences. The ontology serves as the foundational structure for the ML model, facilitating the integration of various comfort factors such as thermal, acoustic, visual comfort, and air quality. Through this integration, the system can process large datasets, predict future comfort levels, and optimize building systems accordingly. Here are the steps for Integration:

- Data Collection:
  Real-time environmental data, such as temperature, humidity, air quality, noise levels, and lighting conditions, are collected through sensors placed within the building. Additionally, occupant preferences regarding comfort levels are gathered through feedback mechanisms (e.g., user interfaces or mobile applications). This data serves as the primary input for training and continuously refining the ML model.
- Training the ML Model:
  Using both historical data (e.g., building performance data) and real-time data, the ML model is trained to understand how different environmental parameters influence indoor comfort. Various algorithms, such as Random Forest, Artificial Neural Networks (ANN), or Support Vector Machines (SVM), can be used to model the complex relationships between comfort factors and environmental conditions. The model learns to correlate these factors with the comfort preferences of building occupants and adjusts settings accordingly.

- Prediction and Adjustment:
  The ML model continuously predicts future comfort conditions based on current environmental data and historical patterns. Using these predictions, the system proactively adjusts building systems (e.g., HVAC, lighting, and ventilation) to maintain optimal comfort levels. For example, if the system predicts a drop in indoor temperature, it can adjust the heating system in advance to prevent discomfort. Similarly, lighting levels can be dynamically adjusted based on predicted daylight availability and occupant preferences.
- Feedback Loop:
  A feedback loop is established where occupants provide feedback on their comfort levels. This feedback, gathered through user interfaces or mobile applications, is fed back into the ML model. Over time, the model refines its predictions based on this feedback, becoming more accurate and responsive to individual comfort preferences. The continuous learning process enables the system to adapt to changing environmental conditions and evolving occupant needs.

By integrating machine learning into the comfort assessment framework, the system becomes a more adaptive and responsive tool for managing indoor environments. Unlike traditional systems, which rely on fixed rules and manual adjustments, this approach allows for continuous, real-time optimization. The result is a more personalized and efficient comfort management system that improves both occupant satisfaction and energy efficiency.

6.9 Summary of CBContology

This chapter presents the development of the Comprehensive Building Comfort Ontology (CBContology), an ontology-based framework designed to assess and enhance indoor comfort by integrating machine learning and user-defined preferences. This chapter focusing on its ability to dynamically calculate comfort levels and provide actionable recommendations for improving building environments.

The CBContology framework was successfully implemented and integrated with machine learning models to provide a robust system for assessing and optimizing indoor comfort. The system was able to model and calculate comprehensive comfort scores (CBC) based on four key comfort dimensions: thermal comfort, acoustic comfort, visual comfort, and air quality. The ontology enabled the dynamic integration of real-time data, user preferences, and environmental conditions to make comfort adjustments. The system functioned as intended, with SWRL rules facilitating the calculation of CBC scores by weighting each comfort factor based on predefined preferences or user input. The integration of machine learning algorithms, such as Random Forest and Artificial Neural Networks (ANN), further enhanced the system's capacity to predict future comfort conditions, making it adaptable and forward-looking.

The implementation of CBContology proved to be a significant advancement over traditional comfort assessment methods. Several key benefits were observed:

- Customizability: One of the major strengths of the CBContology framework is its ability to incorporate user-defined preferences. This feature allows users to specify their individual comfort priorities, such as placing more weight on thermal comfort over acoustic comfort, leading to a personalized and adaptive comfort assessment.
- Real-Time Adjustments: The integration with machine learning models allowed the system to make real-time adjustments based on environmental changes and user feedback. This dynamic aspect made CBContology highly responsive to changing conditions, resulting in a more precise and up-to-date evaluation of comfort levels.
- Comprehensive Comfort Assessment: Unlike traditional methods that tend to focus on one or two comfort factors (such as thermal comfort), CBContology provided a holistic evaluation by incorporating multiple comfort dimensions. This comprehensive approach resulted in more balanced and complete assessments.
- Energy Efficiency: By enabling real-time predictions and dynamic adjustments, the framework also contributed to more energy-efficient building operations. Adjustments to systems like

HVAC, lighting, and ventilation were more targeted, optimizing resource use without compromising comfort.

CBContology introduced a shift in how comfort is assessed and managed in building environments. Traditional methods of comfort evaluation typically rely on standardized guidelines and static models, which do not account for individual preferences or real-time data. CBContology's introduction of a user-driven, ontology-based system allowed for customized, dynamic comfort assessments that align more closely with real-world occupant needs and preferences. This difference is particularly notable in the way CBContology handles user preferences. By leveraging AHP (Analytic Hierarchy Process), the system allowed users to rank their comfort priorities, such as giving more importance to air quality in office settings or thermal comfort in residential spaces. This customization added a human-centric aspect to comfort assessments, making the process more relevant and personalized. The integration of machine learning also made a significant impact, as it allowed the system to predict comfort conditions and optimize energy usage proactively. This predictive capability helped reduce energy waste and enhanced building performance over time.

Several aspects of CBContology worked particularly well:

- Machine Learning Integration: The use of machine learning, particularly Random Forest and Artificial Neural Networks (ANN), enabled highly accurate predictions of comfort levels based on real-time data. The high correlation coefficients observed in the test results validated the effectiveness of these models in predicting comfort parameters.
- SWRL Rule Formulation: The application of SWRL rules to compute comprehensive comfort scores was a success. The rules were able to accurately calculate the CBC based on the weighted contributions of thermal comfort, acoustic comfort, visual comfort, and air quality. This dynamic calculation provided real-time updates on comfort conditions, allowing the system to adjust as needed.
- User Preference Customization: The AHP-based customization process worked seamlessly, enabling users to define their preferences and see those preferences reflected in the comfort assessments. This feature made the system flexible and adaptable, adding significant value compared to rigid, one-size-fits-all approaches.
- Ontology Management: The use of Protégé as an ontology management tool proved highly effective. It facilitated the creation and management of a complex ontology that could evolve as new data or rules were introduced. This ensured that the CBContology system was scalable and capable of handling expansions or modifications as needed.

In conclusion, CBContology proved to be an effective and innovative framework for assessing and optimizing indoor comfort. Its strengths lie in its customizability, real-time dynamic adjustments, and the integration of machine learning to predict and adapt comfort conditions. The use of an ontology-based approach provided the system with flexibility, allowing it to handle complex comfort scenarios and account for personalized user preferences. The success of CBContology suggests that it is a viable solution for modern smart buildings, particularly in contexts where personalization and efficiency are key priorities. The development of the Room Comfort Assessment and Design Suggestion Ontology represents a significant advancement in the field of environmental comfort analysis. By leveraging semantic technologies and SWRL, this research offers a robust framework for the nuanced assessment of room comfort and the generation of targeted design suggestions. Future work will focus on expanding the ontology to include more comfort factors and integrating machine learning techniques for predictive comfort assessment.

7 Conclusion, limitations and Future Work

The study began by exploring the current state and recent developments in BIM, machine learning, and the multi-dimensional nature of building comfort. Through a structured methodological approach, this research laid the foundation for applying advanced technologies to the domain, detailing every phase from data generation to model validation. This dissertation has embarked on a comprehensive journey through the integration of Building Information Modelling (BIM) and machine learning to enhance building comfort and energy efficiency. It commenced with an exploration of the current state and developments in BIM, machine learning, and the multifaceted aspects of building comfort. The methodology chapter laid down a structured approach to applying these technologies, emphasizing a robust scientific design and detailing each step from data generation to model validation.

The establishment of a comprehensive comfort framework represents a significant stride in the field. By adopting a quantitative approach and delving into the four dimensions of comfort, this research has normalized and standardized a complex array of comfort variables into a single, comprehensive metric. This unified metric not only simplifies the assessment of building comfort but also enhances the ability to make informed decisions, improving interaction and collaboration efficiency among architects, engineers, and other stakeholders.

The incorporation of machine learning has marked a transformative shift from traditional modelling methods, offering a more efficient and accurate prediction of building performance. It has addressed the traditional time-consuming and often biased energy simulations, paving the way for automated, intelligent control over the building's life cycle. The research highlighted the potential of machine learning to revolutionize building design and management practices by providing a holistic approach to designing for energy efficiency and occupant comfort. However, the journey does not conclude here. This research has identified several avenues for future exploration, including refining the machine learning models for greater accuracy and exploring additional data sources for a more comprehensive analysis. The balance between energy consumption and comfort remains a critical area of ongoing research, necessitating continuous improvement and innovation in both technological and design strategies.

In closing, this dissertation contributes to the existing body of knowledge by providing a detailed account of integrating BIM and machine learning and ontology to achieve a holistic understanding and enhancement of building comfort and energy efficiency. It serves as a stepping stone for further research and development in the field, encouraging continued exploration and application of these technologies to create more sustainable, comfortable, and efficient buildings for future generations.

7.1 Key Contributions and Findings

One of the major contributions of this research is the establishment of a comprehensive comfort framework, which represents a significant advance in the field of building design and management. By adopting a quantitative approach and incorporating four key dimensions of comfort—thermal, acoustic, visual, and indoor air quality—this framework normalizes a complex set of comfort variables into a unified metric. This metric simplifies the assessment of building comfort, allowing for more informed decision-making by architects, engineers, and other stakeholders. It enables better communication and more effective collaboration, ensuring that comfort is no longer a secondary consideration but a core element in the design and operation of buildings.

The integration of machine learning (ML) into this framework has revolutionized traditional building performance predictions, offering an efficient, accurate, and dynamic alternative to conventional simulation models. Machine learning addressed the time-consuming and often biased nature of traditional energy simulations, automating much of the process and ensuring real-time predictions. This study highlighted the potential for machine learning to not only automate control systems across a building's lifecycle but also optimize comfort and energy efficiency dynamically. The integration of BIM-based ML engines provided a pathway to achieving holistic control over energy consumption and occupant comfort.

Moreover, the introduction of ontology in the latter stages of the research enabled a knowledge-based approach to managing building comfort. The ontology facilitated a more structured, customizable, and flexible approach to comfort assessment, allowing user preferences to be integrated into the comfort models, making them more adaptive to individual needs. The CBContology framework provided a robust system for dynamic adjustments based on real-time data, enhancing the practical application of these theoretical models in live building environments.

7.2 Limitations

Despite the successful outcomes of this study, there are several limitations that were encountered:

Data Limitations: While the use of mock-up data is highly effective in the early design phase for evaluating and refining architectural plans, it presents limitations when applied to the building's entire lifecycle or for improving comfort in existing structures. In such cases, incorporating real-world data could further enhance the robustness and accuracy of the models, allowing for more precise predictions and adjustments based on actual conditions and occupant experiences.

Model Complexity: As more variables are incorporated into the comfort framework (e.g., building orientation, shading devices, material properties), the complexity of the models increases, which may lead to challenges in computation and require more advanced hardware and algorithms for processing.

Generalizability: The study primarily focused on residential buildings in the UK. While the methodology can be applied to other types of buildings or regions, the specific comfort metrics and energy considerations may vary, requiring additional adaptation.

User Preference Implementation: Although the Analytic Hierarchy Process (AHP) was used to incorporate user preferences into the model, the process requires further refinement to handle more complex scenarios where multiple users or conflicting preferences exist within the same building.

7.3 Future Work

The integration of BIM, machine learning, and ontology provides a solid foundation, but there is much room for further research and development:

Refining Machine Learning Models: Future studies should focus on refining the machine learning algorithms by integrating more real-time data and expanding the data sources used for training. This would increase the accuracy and generalizability of the comfort predictions.

Expanding the Comfort Framework: Future research could expand the comfort framework by including additional comfort dimensions, such as psychological comfort or spatial comfort, to provide a more holistic assessment of building environments.

Exploring New Data Sources: Incorporating data from smart devices, IoT sensors, and wearable technology could improve the responsiveness of the system, allowing for real-time adaptations to occupant comfort based on immediate feedback.

Energy and Comfort Balance: The balance between energy consumption and comfort remains a critical area of future research. With increasing focus on sustainability and climate resilience, studies should explore more energy-efficient methods of maintaining comfort without compromising occupant well-being.

Real-World Implementation and Testing: A significant next step would be to test the CBContology framework in real-world environments, applying the models to existing or newly constructed buildings to assess their effectiveness in practice. Such tests would provide invaluable feedback on the framework's adaptability and efficiency in diverse settings.

Enhanced User Interaction: Further work is needed to refine the user interaction process, enabling multi-user preferences to be incorporated seamlessly into the comfort models. This will be particularly important in commercial or public buildings, where diverse occupant profiles are present.

7.4 Conclusion

In conclusion, this dissertation has successfully demonstrated the potential of integrating BIM, machine learning, and ontology to enhance building comfort and energy efficiency. The research has provided a structured, adaptable framework that can dynamically assess and improve comfort based on real-time data, user preferences, and environmental conditions. While there are limitations, the potential for this approach to revolutionize building design and management practices is clear.

As the demand for sustainable and human-centred environments grows, this work provides a strong foundation for future research, encouraging further exploration into advanced technologies for creating comfortable, efficient, and adaptive buildings that meet the evolving needs of occupants and society as a whole.

Appendix 1

This appendix demonstrates the implementation of two machine learning models—Linear Regression (LR) and Support Vector Regression (SVR)—using Python's scikit-learn library in Case Study 1. Below is an overview of the code for each method:

LR methods

```
# read CSVdata
import csv
import numpy as np
filename = 'C:\\Users\\53226\\Desktop\\sample1_data.csv'
raw_data = open (filename, 'rt')
reader = csv.reader(raw_data, delimiter=',', quoting=csv.QUOTE_NONE)
x = list(reader)
data =np.array(x).astype('float')
print(data.shape)

# data processing
d = data
X = d[:,:8]
y = d[:,8]
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# fitting a linear regression model
from sklearn import linear_model
from sklearn.linear_model import LinearRegression
reg = linear_model.LinearRegression()
reg.fit(X_train, y_train)
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
            normalize=False)
s = reg.score(X_test, y_test)
print(s)
```

```
y_pred = reg.predict(X_test)

# plotting the train model
import matplotlib.pyplot as plt
plt.scatter(y_test, y_pred)
plt.plot([y.min(),y.max()],[y.min(),y.max()])
```

SVR methods

```
# read CSVdata
import csv
import numpy as np
filename = 'C:\\Users\\53226\\Desktop\\sample1_data.csv'
raw_data = open(filename, 'rt')
reader = csv.reader(raw_data, delimiter=',', quoting=csv.QUOTE_NONE)
x = list(reader)
data =np.array(x).astype('float')
print(data.shape)

# data processing
d = data
X = d[:,:8]
y = d[:,8]
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# fitting a polynomial SVR model
```

```
from sklearn.svm import SVR

poly_svm = SVR(kernel = "rbf", C = 1.0)

poly_svm.fit(X_train, y_train)

SVR(C=1.0, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma='auto', kernel='rbf',
max_iter=-1, shrinking=True, tol=0.001, verbose=False)

s = poly_svm.score(X_test,y_test)

print(s)

y_pred = poly_svm.predict(X_test)
```

Appendix 2

Appendix 2 contains the dataset used in Case Study, which is stored in an Excel file accessible via a hyperlink below. This dataset was used for training the machine learning models, which includes 8760 instances.

Date/Time	Environment:Site Outdoor Air Drybulb Temperature [C](Hourly)	Environment:Site Outdoor Air Humidity Ratio [kgWater/kgDryAir](Hourly)	Environment:Site Wind Speed [m/s](Hourly)	Environment:Site Diffuse Solar Radiation Rate per Area [W/m2](Hourly)	Environment:Site Direct Solar Radiation Rate per Area [W/m2](Hourly)	Environment:Site Solar Azimuth Angle [deg](Hourly)	Environment:Site Solar Altitude Angle [deg](Hourly)	3145:Zone Windows Total Transmitted Solar Radiation Rate [W](Hourly)
1	8.525	0.006649	5.675	0	0	13.73314	-60.926	0
2	8.325	0.006496	5	0	0	39.31949	-56.7506	0
3	8.55	0.006536	5.8	0	0	59.28305	-49.6835	0
4	8.6	0.006679	6.575	0	0	74.78922	-41.1099	0
5	8.825	0.00705	8.2	0	0	87.64132	-31.9081	0
6	9.5	0.007469	6.45	0	0	99.14265	-22.6113	0
7	10.075	0.007538	5.25	0	0	110.1509	-13.5984	0
8	10.125	0.007731	4.35	0	0	121.2574	-5.20262	0
9	10.625	0.007902	4.475	14.25	11	132.8861	2.23384	2.509658
10	10.725	0.007899	4.225	52.75	90.75	145.3104	8.338182	10.02342
11	10.7	0.00814	3.35	81.75	224.75	158.6091	12.72221	16.34238
12	10.625	0.007939	3.1	104.75	190.75	172.6063	15.04087	19.74009
13	10.675	0.007755	3.475	116	128	186.8789	15.08278	20.48618
14	11.15	0.007976	5.55	87.5	57.75	200.8946	12.84392	15.04453
15	11.225	0.007829	5	46	0	214.2233	8.52915	7.800996
16	10.9	0.00766	3.85	12	0	226.6793	2.48033	2.071693
17	10.125	0.007712	2.85	0	0	238.3329	-4.91461	0
18	10.2	0.007592	2.225	0	0	249.4518	-13.2818	0
19	10.375	0.007579	3.975	0	0	260.4545	-22.2781	0
20	10.475	0.007688	6.175	0	0	271.9251	-31.5707	0
21	10.575	0.007692	5.125	0	0	284.7089	-40.7839	0
22	10.525	0.007516	4.225	0	0	300.0898	-49.3927	0
23	10.425	0.007113	5.3	0	0	319.86	-56.5356	0
24	9.95	0.006847	4.5	0	0	345.2575	-60.8434	0
25	9.875	0.006948	5.3	0	0	13.49032	-60.8618	0
26	9.975	0.006682	7.2	0	0	39.07933	-56.7217	0
27	9.4	0.006547	6.35	0	0	59.07434	-49.6783	0
28	8.675	0.006584	4.55	0	0	74.61008	-41.1171	0

References

Abouelaziz, I., & Jouane, Y. (2023, December). Photogrammetry and deep learning for energy production prediction and building-integrated photovoltaics decarbonization. In *Building Simulation* (pp. 1-17). Beijing: Tsinghua University Press.

Acharya, D., Khoshelham, K., & Winter, S. (2019). BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *150*(March), 245–258. https://doi.org/10.1016/j.isprsjprs.2019.02.020

Al-Alawi, S. M. & Al-Hinai, H. A. (1998) 'An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation', *Renewable Energy*, 14(1–4), pp. 199–204. doi: 10.1016/S0960-1481(98)00068-8.

Al-Homoud, M. S. (2009) 'Envelope thermal design optimization of buildings with intermittent occupancy', *Journal of Building Physics*, 33(1), pp. 65–82. doi: 10.1177/1744259109102799.

Alpaydin, E., 2014. *Introduction to machine learning*. MIT press.

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*(September 2017), 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Arntz, M., Ben Yahmed, S., & Berlingieri, F. (2020). Working from home and COVID-19: The chances and risks for gender gaps. *Intereconomics*, *55*(6), 381-386.

Aydinalp, M., Ugursal, V. I. and Fung, A. S. (2004) 'Modelling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks', *Applied Energy*, 79(2), pp. 159–178. doi: 10.1016/j.apenergy.2003.12.006.

Aydinalp-Koksal, M. and Ugursal, V. I. (2008) 'Comparison of neural network, conditional demand analysis, and engineering approaches for modelling end-use energy consumption in the residential sector', *Applied Energy*, 85(4), pp. 271–296. doi: 10.1016/j.apenergy.2006.09.012.

Jayan, B. (2016). *Real-time Multi-scale Smart Energy Management and Optimisation (REMO) for buildings and their district* (Doctoral dissertation, Cardiff University).

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H.A., Alaka H. A., Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, *30*(3), 500–521.

Bloch, T., & Sacks, R. (2018a). Comparing machine learning and rule-based inferencing for semantic enrichment of BIM models. *Automation in Construction*, *91*(March), 256–272.

Bogen, A. C., Rashid, M., East, E. W., & Ross, J. (2013). Evaluating a data clustering approach for life-cycle facility control. *Journal of Information Technology in Construction*, *18*, 99–118.

Borrmann, A., König, M., Koch, C., & Beetz, J. (2018). *Building information modelling: Why? what? how?* (pp. 1-24). Springer International Publishing.

Bull, R., Chang, N., & Fleming, P. (2012). The use of building energy certificates to reduce energy consumption in European public buildings. *Energy and Buildings*, *50*, 103–110. https://doi.org/10.1016/j.enbuild.2012.03.032

Carnot, N., Koen, V., & Tissot, B. (2016). Modelling Behaviour. *Economic Forecasting*, (buildingSMART 2014), 103–132. https://doi.org/10.1057/9780230005815_5

Carrara, G., Fioravanti, A., Loffreda, G., & Trento, A. (2009). An Ontology-Based Knowledge Representation Model for Cross-disciplinary Building Design-A General Template. *Computation: the new Realm of Architectural Design* (pp. 367-373). Cenkler Printing.

Chen, S. Y. (2019). Enhancing Validity of Green Building Information Modelling with Artificial-neural-network-supervised Learning--Taking Construction of Adaptive Building Envelope Based on Daylight Simulation as an Example. *Sensors & Materials*, *31*.

Dan, T. X., & Phuc, P. N. K. (2018). Application of Machine Learning in Forecasting Energy Usage of Building Design. *Proceedings 2018 4th International Conference on Green Technology and Sustainable Development, GTSD 2018*, 53–59. https://doi.org/10.1109/GTSD.2018.8595595

Dhillon, R. K., Jethwa, M., & Rai, H. S. (2014). Extracting building data from BIM with IFC. *International Journal on Recent Trends in Engineering & Technology*, *11*(2), 202.

Dounis, A. I. (2010) ‘Artificial intelligence for energy conservation in buildings’, *Advances in Building Energy Research*, 4(1), pp. 267–299. doi: 10.3763/aber.2009.0408.

Eadie, R., Browne, M., Odeyinka, H., McKeown, C., & McNiff, S. (2013). BIM implementation throughout the UK construction project lifecycle: An analysis. *Automation in Construction*, *36*, 145–151.

Eiter, T., Ianni, G., Krennwallner, T., & Polleres, A. (2008). Rules and ontologies for the semantic web. *Reasoning Web: 4th International Summer School 2008, Venice, Italy, September 7-11, 2008, Tutorial Lectures*, 1-53.

Eleftheriadis, S., Mumovic, D., & Greening, P. (2017). Life cycle energy efficiency in building structures: A review of current developments and future outlooks based on BIM capabilities. *Renewable and Sustainable Energy Reviews*, *67*, 811-825.

Eurostat. (2020). LFS series (lfsa_ehomp, lfst_hhnhtych), Statistical office of the European Union. https://ec.europa.eu/eurostat/databrowser/view/LFSA_EHOMP/default/line?lang=en

Fanger, P. O. (1992). Efficient ventilation for human comfort.

Ganesh, G. A., Sinha, S. L., Verma, T. N., & Dewangan, S. K. (2021). Investigation of indoor environment quality and factors affecting human comfort: A critical review. *Building and Environment*, *204*, 108146.

Genesereth, M. R., & Nilsson, N. J. (2012). *Logical foundations of artificial intelligence*. Morgan Kaufmann.

Golparvar-Fard, M., Bohn, J., Teizer, J., Savarese, S., & Peña-Mora, F. (2011). Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques. *Automation in construction*, *20*(8), 1143-1155.

Golparvar-Fard, M., Peña-Mora, F., & Savarese, S. (2012). Automated Progress Monitoring Using Unordered Daily Construction Photographs and IFC-Based Building Information Models. *Journal of Computing in Civil Engineering*, *29*(1), 04014025. https://doi.org/10.1061/(asce)cp.1943-5487.0000205

Hamdy, M. et al. (2011). Impact of adaptive thermal comfort criteria on building energy use and cooling equipment size using a multi-objective optimization scheme. Energy and Buildings 43(9), pp. 2055–2067.

Han, K., & Golparvar-Fard, M. (2017). Crowdsourcing BIM-guided collection of construction material library from site photologs. *Visualization in Engineering*, *5*(1). https://doi.org/10.1186/s40327-017-0052-3

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, 485-585.

Hernández, L., Baladrón, C., Aguiar, J. M., Calavia, L., Carro, B., Sánchez-Esguevillas, A., Fernández A., & Lloret, J. (2014). Artificial neural network for short-term load forecasting in distribution systems. *Energies*, *7*(3), 1576-1598.

Hopgood, A. A. (2021). *Intelligent systems for engineers and scientists: a practical guide to artificial intelligence*. CRC press.

Hou, S. (2015). *An ontology-based holistic approach for multi-objective sustainable structural design* (Doctoral dissertation, Cardiff University).

Hu, Y., & Castro-Lacouture, D. (2018). Clash Relevance Prediction Based on Machine Learning. *Journal of Computing in Civil Engineering*, *33*(2), 04018060. https://doi.org/10.1061/(asce)cp.1943-5487.0000810

Hunt, V. D. (2012). *Artificial intelligence & expert systems sourcebook*. Springer Science & Business Media.

Idowu, S. *et al.* (2014) 'Forecasting heat load for smart district heating systems: A machine learning approach', *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pp. 554–559. doi: 10.1109/SmartGridComm.2014.7007705.

ISO. (2008). Energy performance of buildings: Calculation of energy use for space heating and cooling (ISO 13790:2008). Geneva, Switzerland: ISO.

Jain AK, MurtyMN, Flynn P (1999) Data clustering: a review. ACMComput Surveys 31(3):264–323

Jung, N., & Lee, G. (2019). Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning. *Advanced Engineering Informatics*, *41*(April), 100917. https://doi.org/10.1016/j.aei.2019.04.007

Kalogirou, S. A. (1999). Applications of artificial neural networks in energy systems. *Energy Conversion and Management*, *40*(10), 1073-1087.

Kalogirou, S. A., & Bojic, M. (2000). Artificial neural networks for the prediction of the energy consumption of a passive solar building. *Energy*, *25*(5), 479-491.

Kalogirou, S. A., Eftekhari, M. M., & Pinnock, D. J. (1999). Prediction of air flow in a single-sized naturally ventilated test room using artificial neural networks.

Kalogirou, S. A., Michaelides, S., & Tymvios, F. S. (2002). Prediction of maximum solar radiation using artificial neural networks. In *World Renewable Energy Congress VII*.1–5. Available at: http://ktisis.cut.ac.cy/handle/10488/877

Kalogirou, S. A., Neocleous, C. C., & Schizas, C. N. (1997, November). Building heating load estimation using artificial neural networks. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques* (Vol. 8, p. 14).

Kalogirou, S. A., & Panteliou, S. (2000). Thermosiphon solar domestic water heating systems: long-term performance prediction using artificial neural networks. *Solar Energy*, *69*(2), 163-174.

Kamel, E., & Memari, A. M. (2019). Review of BIM's application in energy simulation: Tools, issues, and solutions. *Automation in Construction*, *97*(October 2018), 164–180. https://doi.org/10.1016/j.autcon.2018.11.008

Kandananond, K. (2011) 'Forecasting electricity demand in Thailand with an artificial neural network approach', *Energies*, 4(8), pp. 1246–1257. doi: 10.3390/en4081246.

Karan, E., & Asadi, S. (2019). Intelligent designer: A computational approach to automating design of windows in buildings. *Automation in Construction*, *102*(February), 160–169. https://doi.org/10.1016/j.autcon.2019.02.019

Kemmoku, Y. *et al.* (1999) 'Daily insolation forecasting using a multi-stage neural network', *Solar Energy*, 66(3), pp. 193–199. doi: 10.1016/S0038-092X (99)00017-1.

Kim, J. B., Jeong, W., Clayton, M. J., Haberl, J. S., & Yan, W. (2015). Developing a physical BIM library for building thermal energy simulation. *Automation in Construction*, *50*(C), 16–28. https://doi.org/10.1016/j.autcon.2014.10.011

Kittler, R., Kocifaj, M., & Darula, S. (2011). Daylight science and daylighting technology. Springer Science & Business Media.

Kodratoff, Y. (2014). Introduction to machine learning. *Elsevier*.

Kohonen, T. (1988) 'An introduction to neural computing', *Neural Networks*, 1(1), pp. 3–16. doi: 10.1016/0893-6080(88)90020-2.

Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E., (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), pp.159-190.

Krarti, M. (2003) 'An Overview of Artificial Intelligence-Based Methods for Building Energy Systems', *Journal of Solar Energy Engineering*, 125(3), p. 331. doi: 10.1115/1.1592186.

Krygiel, E., Nies, B. and McDowell, S. (2008) *Green BIM : Successful Sustainable Design with Building Information Modeling.* John Wiley & Sons.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Nature Publishing Group, 521(7553), pp. 436–444. doi: 10.1038/nature14539.

Lin, J. R. *et al.* (2016) 'A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM', *Computer-Aided Civil and Infrastructure Engineering*, 31(1), pp. 18–33. doi: 10.1111/mice.12151.

Lin, K.-L., Jan, M.-Y. and Liao, C.-S. (2017) 'Energy Consumption Analysis for Concrete Residences—A Baseline Study in Taiwan', *Sustainability*, 9(2), p. 257. doi: 10.3390/su9020257.

López, G., Batlles, F. J. and Tovar-Pescador, J. (2005) 'Selection of input parameters to model direct solar irradiance by using artificial neural networks', *Energy*, 30(9 SPEC. ISS.), pp. 1675–1684. doi: 10.1016/j.energy.2004.04.035.

Lork, C., Choudhary, V., Hassan, N. U., Tushar, W., Yuen, C., Ng, B. K. K., ... & Liu, X. (2019). An ontology-based framework for building energy management with IoT. Electronics, 8(5), 485.

Marmaras, C. *et al.* (2017) 'Predicting the energy demand of buildings during triad peaks in GB', *Energy and Buildings*. Elsevier B.V., 141, pp. 262–273. doi: 10.1016/j.enbuild.2017.02.046.

Mihalakakou, G., Santamouris, M. and Tsangrassoulis, A. (2002) 'On the energy consumption in residential buildings', *Energy and Buildings*, 34(7), pp. 727–736. doi: 10.1016/S0378-7788(01)00137-2.

Mitchell, T. (1997). *Machine learning.* McGraw Hill, New York, NY.

Mohd-Nor, M. F. I. and Grant, M. P. (2014) 'Building information modelling (BIM) in the malaysian architecture industry', *WSEAS Transactions on Environment and Development*, 10, pp. 264–273. doi: 10.1016/j.buildenv.2006.10.027.

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.

Pacheco-Torgal, F., Jalali, S., & Fucic, A. (Eds.). (2012). *Toxicity of building materials*. Elsevier.

Pfafferott, J., Herkel, S. and Wapler, J. (2005) 'Thermal building behaviour in summer: Long-term data evaluation using simplified models', *Energy and Buildings*, 37(8), pp. 844–852. doi: 10.1016/j.enbuild.2004.11.007.

Poux, F. (2019). *The smart point cloud: Structuring 3D intelligent point data* (Doctoral dissertation, University of Liège).

Prior, J. (1993). Building research establishment environmental assessment method. *BREEAM) Version*, *1*, 93.

Qin, F., Li, L., Gao, S., Yang, X., & Chen, X. (2014). A deep learning approach to the classification of 3D CAD models. *Journal of Zhejiang University SCIENCE C*, *15*(2), 91–106. https://doi.org/10.1631/jzus.c1300185

Reddy, K. S. and Ranjan, M. (2003) 'Solar resource estimation using artificial neural networks and comparison with other correlation models', *Energy Conversion and Management*, 44(15), pp. 2519–2530. doi: 10.1016/S0196-8904(03)00009-8.

Rehman, S., & Mohandes, M. (2008). Artificial neural network estimation of global solar radiation using air temperature and relative humidity. *Energy policy*, *36*(2), 571-576.

Ryu, M. (2020). *Machine learning-based classification system for building information models*. (Mater thsis, Aalto University) Available at: https://urn.fi/URN:NBN:fi:aalto-202109149135

Schmidhuber, J. (2015) 'Deep learning in neural networks: An overview', *Neural Networks*. Pergamon, 61, pp. 85–117. doi: 10.1016/J.NEUNET.2014.09.003.

Schreiber, G. (2000). Knowledge engineering and management: the CommonKADS methodology. MIT press.

Serra, R. (1998). Daylighting. *Renewable and sustainable energy reviews*, *2*(1-2), 115-155.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Tang, P., Huber, D., Akinci, B., Lipman, R., & Lytle, A. (2010). Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in construction*, *19*(7), 829-843.

Teizer, J. (2015). Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Advanced Engineering Informatics*, *29*(2), 225–238. https://doi.org/10.1016/j.aei.2015.03.006

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2017). Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining. *Automation in Construction*, *74*, 39–54. https://doi.org/10.1016/j.autcon.2016.11.001

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, *11*(2), 93-136.

Vardaxis, N. G., Bard, D., & Persson Waye, K. (2018). Review of acoustic comfort evaluation in dwellings—part I: Associations of acoustic field data to subjective responses from building surveys. *Building Acoustics*, *25*(2), 151-170.

Wang, X. *et al.* (1999) 'Gray predicting theory and application of energy consumption of building heat-moisture system', *Building and Environment*, 34(4), pp. 417–420. doi: 10.1016/S0360-1323(98)00037-7.

WCED, S. W. S. (1987). World commission on environment and development. *Our common future*, *17*(1), 1-91.

Weinberger, K. R., Zanobetti, A., Schwartz, J., & Wellenius, G. A. (2018). Effectiveness of National Weather Service heat alerts in preventing mortality in 20 US cities. *Environment international*, *116*, 30-38.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., (2016) *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yang, I. H., Yeo, M. S. and Kim, K. W. (2003) 'Application of artificial neural network to predict the optimal start time for heating system in building', *Energy Conversion and Management*, 44(17), pp. 2791–2809. doi: 10.1016/S0196-8904(03)00044-X.

Yang, J., Shi, Z. K., & Wu, Z. Y. (2016). Towards automatic generation of as-built BIM: 3D building facade modelling and material recognition from images. *International Journal of Automation and Computing*, *13*(4), 338–349. https://doi.org/10.1007/s11633-016-0965-7

Yuce, B., Mourshed, M. and Rezgui, Y. (2017) 'A smart forecasting approach to district energy management', *Energies*, 10(8), pp. 1–22. doi: 10.3390/en10081073.

Zaheer-Uddin, M. and Tudoroiu, N. (2004) 'Neuro-models for discharge air temperature system', Energy Conversion and Management, 45(6), pp. 901–910. doi: 10.1016/j.enconman.2003.08.004.

Li, Q., Ren, P., & Meng, Q. (2010, June). Prediction model of annual energy consumption of residential buildings. In 2010 international conference on advances in energy engineering (pp. 223-226). IEEE.

Zhao, H. (2011). *Artificial intelligence models for large scale buildings energy consumption analysis* (Doctoral dissertation, Ecole Centrale Paris).

Zhao, H. and Magoulès, F. (2010) 'Parallel Support Vector Machines Applied to the Prediction of Multiple Buildings Energy Consumption', *Journal of Algorithms & Computational Technology* Vol, 4(2), p. 231:250.

Zhao, H. and Magoules, F. (2011) 'New parallel support vector regression for predicting building energy consumption', *2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM)*, (3), pp. 14–21. doi: 10.1109/SMDCM.2011.5949289.

Zhao, H. and Magoulès, F. (2012) 'A review on the prediction of building energy consumption', *Renewable and Sustainable Energy Reviews*, 16(6), pp. 3586–3592. doi: 10.1016/j.rser.2012.02.049.

Zmeureanu, R. (2002) 'Prediction of the COP of existing rooftop units using artificial neural network and minimum number of sensors', *Energy*, 27(9), pp. 889–904. doi: 10.1016/S0360-5442(02)00027-0.