AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE: CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

RIMA ALAAEDDINE

UNIVERSITY OF HUDDERSFIELD

AN ESEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE: CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

RIMA ALAAEDDINE

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy in Built Environment

UNIVERSITY OF HUDDERSFIELD

NOVEMBER 2023

DECLARATION

I declare that this thesis entitled *"*an ensemble model for predictive energy performance: closing the gap between actual and predicted energy use in residential buildings *"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Signature	:	.....rima alaaeddine.....
Name	:	RIMA ALAAEDDINE
Date	:	1 NOVEMBER 2023

DEDICATION

I dedicate this work to the memory of my beloved grandfathers, Kheir Eddine and Ammar, whom I miss daily and hope they look down on me with pride.

To my mother, who selflessly sacrificed everything for her children, I hope I have made you proud with this achievement.

To my father, who has always been my unwavering support system, thank you for your love and encouragement.

To my three beautiful nieces, Farah, Christine and Joud, I hope you grow up to achieve all your dreams and goals.

To my siblings, Khair, Mariam, and Karim, you are my backbone and my source of pride.

I also extend my gratitude to my grandma, who is always praying for me, the rest of my relatives, friends I made in the UK that became family to me, and colleagues, including Yre, Mahe, Elham, Omayma, and Oshie, for their constant presence and support throughout my journey.

And finally, I dedicate this achievement to myself, for all the hard work, sacrifices, and dedication that I have put in over the years to reach this point. This degree represents not just my academic achievement but also my personal growth, resilience, and determination. I am proud of myself for overcoming challenges and pushing through difficult times, and for never giving up on my dreams. May this accomplishment be a reminder to myself to always strive for excellence and to never stop learning and growing

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who have supported me in making a meaningful contribution to society. My sincere thanks go to my supervisors, Song Wu, Patricia Tzortzopoulos, Mike Kagiolglou, and Elham Del Zendeh, for their unwavering guidance, encouragement, and inspiration throughout this research. Their help and advice have been invaluable, and I could not have completed this research without them. I also appreciate the support of the University of Huddersfield for providing me with the necessary facilities to carry out this research.

University of Huddersfield, November 2023

RIMA ALAAEDDINE

ABSTRACT

The design stage of a building plays a pivotal role in influencing its life cycle and overall performance. Accurate predictions of a building's performance are crucial for informed decision-making, particularly in terms of energy performance, given the escalating global awareness of climate change and the imperative to enhance energy efficiency in buildings. However, a well-documented energy performance gap persists between actual and predicted energy consumption, primarily attributed to the unpredictable nature of occupant behavior.

Existing methodologies for predicting and simulating occupant behavior in buildings frequently neglect or exclusively concentrate on particular behaviors, resulting in uncertainties in energy performance predictions. Machine learning approaches have exhibited increased accuracy in predicting occupant energy behavior, yet the majority of extant studies focus on specific behavior types rather than investigating the interactions among all contributing factors.

This dissertation delves into the building energy performance gap, with a particular emphasis on the influence of occupants on energy performance. A comprehensive literature review scrutinizes machine learning models employed for predicting occupants' behavior in buildings and assesses their performance. The review uncovers knowledge gaps, as most studies are case-specific and lack a consolidated database to examine diverse behaviors across various building types.

An ensemble model integrating occupant behavior parameters is devised to enhance the accuracy of energy performance predictions in residential buildings. Multiple algorithms are examined, with the selection of algorithms contingent upon evaluation metrics. The ensemble model is validated through a case study that compares actual energy consumption with the predictions of the ensemble model and an EnergyPlus simulation that takes occupant behavior factors into account.

The findings demonstrate that the ensemble model provides considerably more accurate predictions of actual energy consumption compared to the EnergyPlus simulation. This dissertation also addresses the research limitations, including the reusability of the model and the requirement for additional datasets to bolster confidence in the model's applicability across diverse building types and occupant behavior patterns.

In summary, this dissertation presents an ensemble model that endeavors to bridge the gap between actual and predicted energy usage in residential buildings by incorporating occupant behavior parameters, leading to more precise energy performance predictions and promoting superior energy management strategies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
ANN	-	Artificial Neural Networks
ML	-	Machine Learning
OB	-	Occupant Behaviour
DT	-	Decision Tree
EECPM	-	Ensemble Energy Consumption Prediction Model
MMRE	-	Mean Magnitude of Relative Error
MRE	-	Magnitude of Relative Error
MSE	-	Mean Squared Error
MAPE	-	Mean Absolute Percentage Error
RMSE	-	Root Mean Squared Error
MAD	-	Mean Absolute Deviation
SLR	-	Systematic Literature Review
SVR	-	Support Vector Regression
RF	-	Random Forest
NN	-	Neural Network
HVAC	-	Heating, Ventilation, and Air Conditioning
MPC	-	Model Predictive Control
BPS	-	Building Performance Standards
GB	-	Gradient Boost
RT	-	Regression Tree
MLR	-	Multiple Linear Regression
LSTM	-	Long Short-Term Memory
XGB	-	Extreme Gradient Boosting

LIST OF APPENDICES

AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE:
CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

Introduction
Chapter 1

CHAPTER 1

INTRODUCTION

1.1. Overview

Over recent years, energy consumption in residential buildings has experienced a significant increase, with global energy demand projected to rise by 28% by 2040 (IEA, 2017). This trend emphasizes the urgency of innovative solutions to ensure efficient energy use, thereby reducing the detrimental impacts of excessive consumption on the environment and the economy (Zuhaib et al., 2022). The residential sector alone accounts for approximately 29% of total global energy consumption (IEA, 2021), which further emphasizes the importance of addressing this issue.

Building operations significantly impact energy consumption, with the building sector contributing to a sizable portion (about 36%) of global energy use (Nejat et al., 2015). Consequently, it is essential to enhance building energy efficiency through the implementation of precise and reliable energy consumption prediction models (Y. Jin et al., 2021; Olu-Ajayi et al., 2022b). Accurate prediction models facilitate better decision-making regarding energy conservation measures, retrofitting strategies, and policy formulation (Dong et al., 2023; Jami et al., 2021).

Energy is a vital component of economic and social development (Mohamed & Lee, 2006; Song et al., 2023). However, the rising energy demand, limited resources to produce energy, and the need for high-quality energy at an affordable price have made sustainable growth challenging (Kaygusuz, 2012). The recent increase in energy prices in the UK and the cost of living crisis underscore the significance of finding efficient ways to meet the nation's energy needs (Farghali et al., 2023). In this context, achieving energy efficiency in residential buildings becomes a priority, given the sector's considerable contribution to overall energy consumption.

Occupant behavior plays a crucial role in energy demand, emphasizing the importance of utilizing energy efficiently (Harputlugil & de Wilde, 2021; Janda, 2011; Jia et al., 2017). Factors such as occupants' habits, preferences, and lifestyles significantly influence energy consumption patterns in residential buildings (Nia et al., 2022) (Fig 1.1). Accurate energy demand prediction for residential usage can lead to better energy management and conservation strategies, as it enables the identification of energy-saving opportunities and the tailoring of demand response programs (Pallonetto et al., 2020; Qureshi et al., 2011).

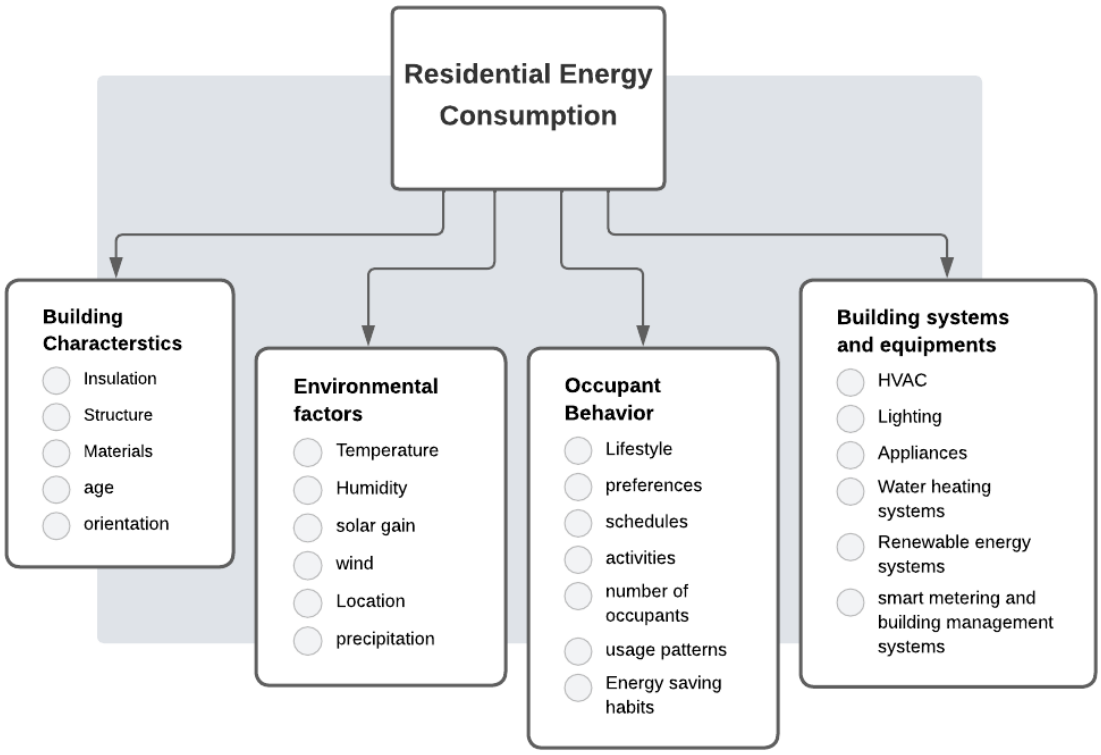


Figure 1-1 Factors influencing residential energy consumption.

Machine learning algorithms have shown promise in energy consumption prediction, accounting for the complex interplay between various factors influencing residential energy use (Mohammadiziazi & Bilec, 2020; Pan & Zhang, 2020). Machine learning approaches offer several advantages over traditional and simulation-based methods in energy consumption prediction. First, machine learning models can learn from large datasets and capture complex relationships between input features and energy consumption, thereby providing more accurate predictions (Bouktif et al., 2018). Second,

machine learning algorithms can adapt to changes in the underlying data distribution, making them suitable for handling non-stationary and noisy data (Padakandla et al., 2020).

The performance of machine learning methods in numerous applications has been widely recognized, including energy consumption forecasting, fault detection and diagnosis, and demand-side management (Alamaniotis & Bargiotas, 2020; Fumo & Rocco, 2015). The use of IoT devices to collect context information about the indoor environment, coupled with various machine learning techniques, allows for the anticipation of future circumstances and energy requirements (Mashal et al., 2021; Khan et al., 2020). This integration fosters more informed decision-making by enabling real-time monitoring, control, and optimization of energy consumption in residential buildings (Zhang et al., 2020; Jin et al., 2020).

Existing building energy simulation tools, such as EnergyPlus and TRNSYS, can provide insights into energy consumption patterns (Crawley et al., 2008; Klein et al., 2010). However, these tools often require expert knowledge and detailed information about building characteristics, which may not always be available (Nouvel et al., 2015). Furthermore, these tools are computationally intensive and may not adequately capture the complexities of occupant behavior and their interactions with building systems (Ascione et al., 2016).

Recent advancements in data collection and processing technologies have paved the way for the development of data-driven models for energy consumption prediction. These models leverage data collected from various sources, such as smart meters, building automation systems, and environmental sensors, to train machine learning algorithms capable of predicting energy consumption patterns in residential buildings (Deb et al., 2020; Sanaullah et al., 2020). The integration of these data sources with machine learning techniques offers a more comprehensive understanding of the factors influencing energy consumption and enables the development of more accurate and robust prediction models.

However, despite the promising results achieved by data-driven models, there is still room for improvement in terms of prediction accuracy and generalizability (Zhao & Magoules, 2012; Li et al., 2014). One of the key challenges is the selection of appropriate feature sets and the development of models that can effectively capture the complex interactions between various factors influencing energy consumption. Another challenge lies in the development of models that can adapt to changing conditions and provide reliable predictions under different scenarios.

To address these challenges, the proposed occupancy behavioral-based machine learning ensemble model will incorporate a diverse set of features, including building characteristics, and occupant behavior data. This comprehensive feature set will enable the model to better capture the complex relationships between various factors and their impact on energy consumption in residential buildings.

Additionally, the proposed model will utilize ensemble learning techniques, which have been proven effective in improving the prediction accuracy and generalizability of machine learning models (Zhang & Qi, 2019; Zhou, 2012). By combining the strengths of Lasso, Ridge, Random forrest, and Extreme Gradient Boost algorithms, the ensemble model aims to achieve better overall performance, even in the presence of noisy and non-stationary data.

The development and evaluation of the proposed model will be carried out using real-world datasets collected from residential buildings, ensuring its applicability and relevance to real-life situations (Li et al., 2018; Todeschi et al., 2020). A thorough comparison with existing models and approaches will be conducted to assess the performance of the proposed model and identify potential areas for improvement.

Furthermore, the findings of this research can inform policymakers and stakeholders involved in the planning and development of sustainable urban environments. By identifying the key factors influencing residential energy consumption and demonstrating the effectiveness of machine learning models

in predicting energy use, this research can provide valuable insights for the design and implementation of energy efficiency policies, building codes, and regulations (Pérez-Lombard et al., 2008; Ürge-Vorsatz et al., 2015).

In conclusion, this thesis aims to develop an innovative ensemble model for predictive energy performance in residential buildings, addressing the pressing need for accurate and reliable energy consumption prediction (Fig 1.2). By integrating occupancy behavioral factors with advanced machine learning techniques, the proposed model seeks to close the gap between actual and predicted energy use, ultimately contributing to the sustainable growth and efficient energy management of residential buildings. The findings of this research have the potential to significantly impact various aspects of energy management, from demand response programs to policy development, ultimately fostering a more sustainable and energy-efficient future.

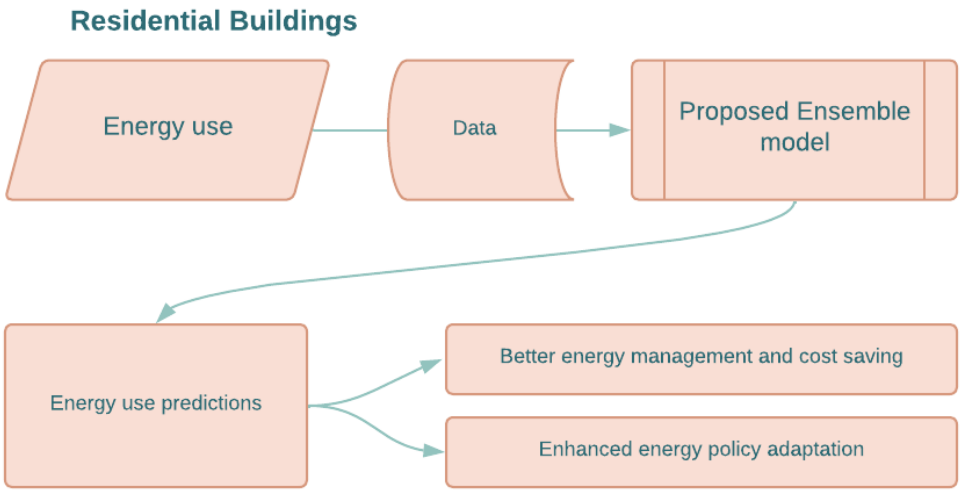


Figure 1-2 Potential applications of the proposed machine learning ensemble model

1.2. Research Background

The prediction models are models that employ data mining and probability to predict outcomes (Witten et al., 2016). Building energy performance prediction models use a set of input parameters to quantify building energy demands (Krstić & Teni, 2017). These prediction models are

commonly used to predict the energy use to identify patterns, changes in energy use, or ensure that energy use meets energy requirements. Building energy performance prediction is an aid in guiding decision making to support building codes, evaluate different design or renovation alternatives, guide occupants and stakeholders (Swan & Ugursal, 2009). In UK, Europe, and worldwide, buildings are bound to reach minimum requirements in terms of energy efficiency as per the Energy Efficiency Requirements in Different Regions table (Table 1.1). Thus, building energy performance prediction became a growing concern area for researchers in their attempt to conserve energy, minimise energy waste and achieve the required energy targets in buildings.

Table 1-1 Energy Efficiency Requirements in Different Regions.

Region	Energy Efficiency Requirements	Reference
UK	• Minimum EPC rating of E for privately rented properties • Minimum EPC rating of E for new tenancies and renewals, and for all private rented properties by 2023 • Ten Point Plan for a Green Industrial Revolution • Higher energy efficiency standards for new homes from 2025	(UK Government, 2018) (UK Government, 2020) (UK Government, 2019) (UK Government, 2023)
Europe	• Various energy efficiency targets for 2020 and 2030 • All new buildings to be nearly zero-energy buildings by end of 2020, and all existing buildings to be renovated to NZEB standard by 2050 • Headline EU energy efficiency target for 2030 of at least 32.5%	(European Commission, 2018) (European Commission, 2021) (European Commission, 2022)
Worldwide	• Efforts to limit temperature increase to 1.5°C, including measures to improve energy efficiency in buildings and other sectors	(UNFCCC, 2019) (IEA TCP, 2021)

Various energy prediction approaches have been developed and utilised depending on different levels of details and input requirements. Literature has revealed and categorized these approaches as black, white and grey box approaches (Borgstein et al., 2016; Foucquier et al., 2013; Koulamas et al., 2017; Krstić & Teni, 2017; Li et al., 2014) shown in Figure 1.3.

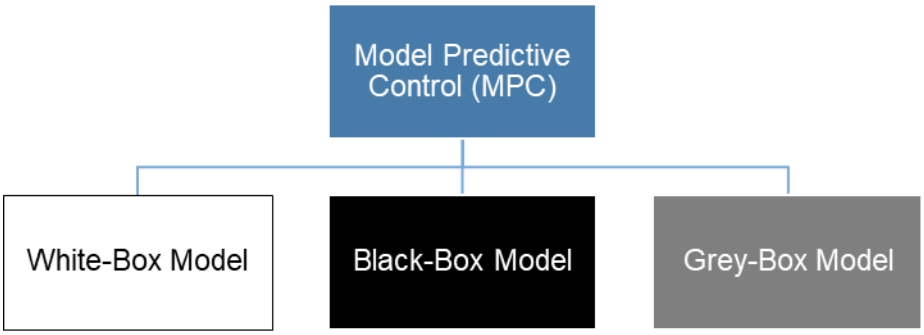


Figure 1-3 MPC approaches based on building energy prediction

White box models, defined as physical models, are calculation-based models that rely on engineering approach to model building components and systems shown in Figure 1.4. These models depend on a high level and details of inputs to ensure the accuracy of the simulation. EnergyPlus (Crawley et al., 2000), IES VE, DOE-2, TRANSYS and ESP-r are the most used white box approaches for energy performance simulations (Castaldo & Pisello, 2018; Zou et al., 2018).

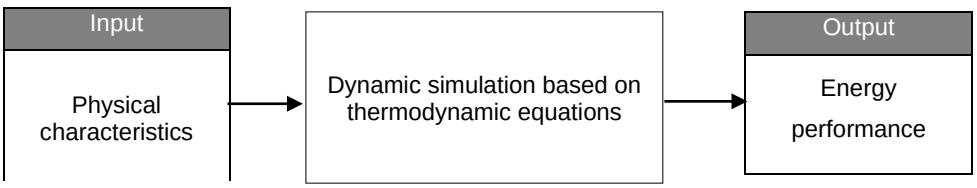


Figure 1-4 White-box energy performance prediction approach

Black box models are data-driven models based on advanced statistical and machine learning approaches shown in Figure 1.5. These models use monitoring/empirical settings to find connection and trends between outputs and inputs, and make deductions and predictions without the required

knowledge of the physical systems (Koulamas et al., 2017; Krstić & Teni, 2017).

Figure 1-5 Black box energy prediction approach

Grey box models are hybrid models that employ statistical and data-driven approach combined with specific knowledge derived from physical building characteristics (Krstić & Teni, 2017). The building energy models accuracy varies according to its purpose, breadth of input parameters, data abundance, and level of accuracy in the data itself. The availability and accuracy of such elements define the reliability of the predictive model to provide appropriate results.

The influence of building occupants on energy usage has been extensively examined, yet there remains a persistent discrepancy between anticipated and actual energy usage in buildings (Delzendeh et al., 2017). This indicates a need for more in-depth research to comprehend the patterns of behavior among occupants. The spectrum of actions by occupants that influence the energy usage of a building includes the operation of appliances, the opening and closing of windows and doors, the utilization of hot water, the adjustment of heating, ventilation, and air conditioning (HVAC) systems such as thermostat settings, the use of lighting, and the manipulation of blinds (Barthelmes, Becchio, et al., 2017). Occupant behavior is not limited to direct interactions with energy-consuming devices; it also includes indirect contributions such as the generation of body heat, which contributes to the internal heat load of a building (Buso et al., 2015). Categorizing the various ways occupants actively and passively contribute to a building's energy profile is crucial for a more nuanced understanding of their overall impact on energy consumption. In their study, Zendeh (2019) offers a delineation of these behaviors in Figure 1.6.

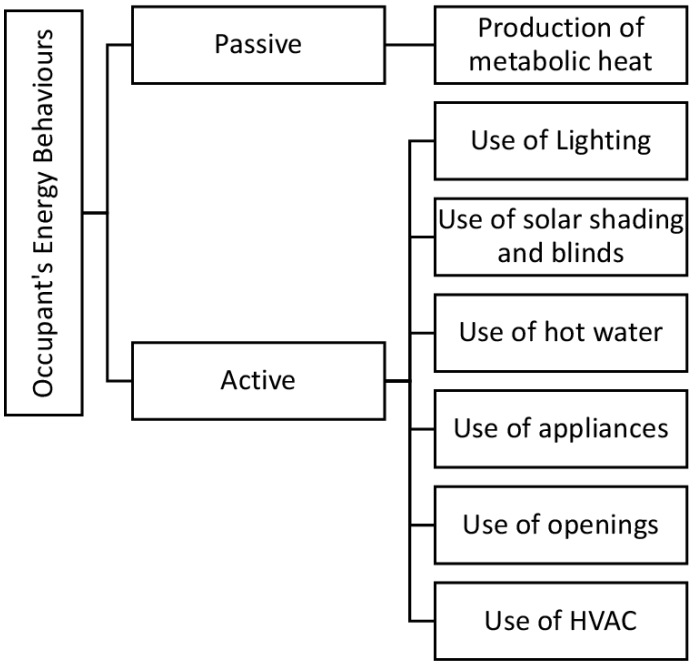


Figure 1-6 Occupant's active and passive energy behaviours (Zendeh, 2019)

A fundamental part for building design and construction is decision making, in which achieving national and international targets, and meeting objectives and requirement is crucial. There is a great emphasis on building energy performance prediction, as well as the enabling of energy efficiency measures amongst the scientific research community as buildings are one of the main energy consuming sectors with an estimate of one third of total energy resources (Paone & Bacher, 2018; Pérez-Lombard et al., 2008). This consumption has been on the rise over the last decade, thus rationalizing the need to minimize building energy performance. To address this concern, more energy efficient building design and operational solutions are put in place (UNEP, 2016), which are projected through energy performance predictions. Energy performance prediction facilitates the exploration of different scenarios and the investigation various solutions to utilize the energy in buildings in the most effective manner (Mehta et al., 2013).

Since the importance of predicting the building energy performance is

established as a mean to achieve energy conservation and promote effective building use Huang et al. (2014), the process of building energy performance prediction is constantly under study to provide more reliable and reinforce building energy optimization. When predicting building energy performance, the outcome is typically determined through building simulation software. Energy simulation software provides the potential to analyze the energy usage patterns and predict the overall energy performance (Huang et al., 2014). Nevertheless, for the model to obtain reliable outcomes, time, skill, and high level of accuracy and details are needed (Buratti et al., 2014).

The prediction relies heavily on the well-informed identification and understanding of all the parameters contributing in the building energy consumption and their integration in the model (Demanuele et al., 2010). Those parameters include building physical and thermal characteristics, climatic and meteorological conditions, building control systems and services, indoor environmental quality requirements, and occupant related inputs (Krstić & Teni, 2017; Yu et al., 2010). The inclusion and the reliable estimation of all these parameters, and the comprehension of their impact on building energy performance results in more accurate energy prediction, which support minimizing the building performance gap, and optimizing energy performance (Yu et al., 2011).

As per definition, the energy performance gap is the gap between actual and predicted energy performance (De Wilde, 2014). One of the most influential factors contributing to this gap is said to be occupant's behaviours and its estimation which, in most cases, is based on unrealistic, oversimplified and unreliable energy-behavioural assumptions shown in Figure 1.7 (Buso et al., 2015; De Wilde, 2014; Haldi & Robinson, 2011; A. C. Menezes et al., 2012). Providing valid energy predictions implies extensive recognition and comprehension of occupant's behaviour in buildings and their impact on energy performance. This currently presents a challenge due to the complicated, stochastic and sophisticated nature of occupant's behaviour (Yan & Malkawi, 2013).

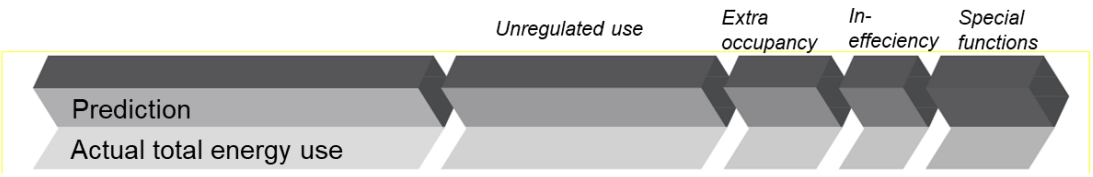


Figure 1-7 Energy performance gap and accompanied uncertainties

Since the impact of occupants behaviour on energy performance is highlighted, more advanced methods that are able to learn from the interaction between the occupant and buildings are needed to predict the building energy performance (Alaaeddine & Wu, 2017; Tam et al., 2018). The energy used in a residential building is influenced by climatic, social, economic and cultural context is shown in Figure 1.8.

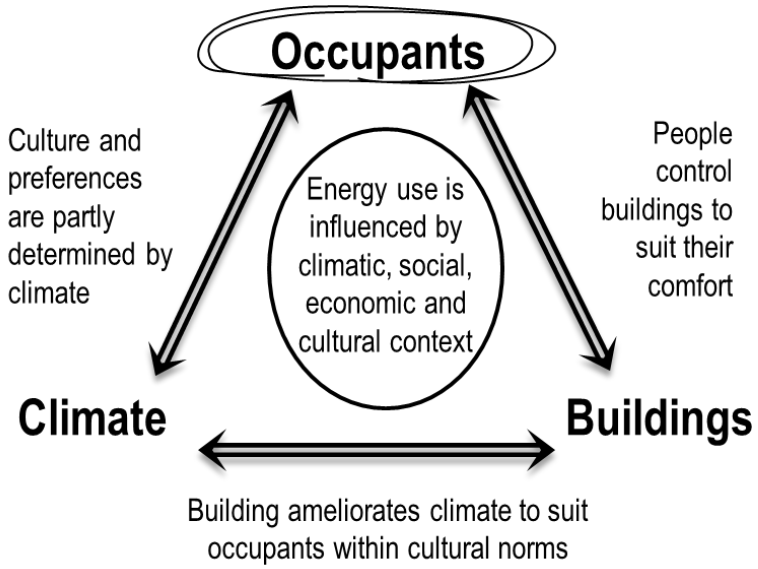


Figure 1-8 Influences on energy use

Numerous attempts to incorporate the impact of occupant's behavioural parameters in simulation models have been made in order to provide more accuracy and reliability (Jang & Kang, 2016; Wang & Ding, 2015; Yu et al., 2011; Zhao & Magoulès, 2012). Those attempts proved that building energy performance calculation has a nonlinear dependency on various number of external and internal variables with high level dimensional data (Huang et al., 2014). The machine learning approaches have been widely studied as prediction approaches for their ability to map nonlinear dependencies, and deal with high dimensional data (Okujeni et al., 2014). Machine learning

approaches are capable to deliver otherwise challenging tasks, such as classification of behavioural patterns, and providing insights and prediction from complex datasets. Therefore, machine learning approaches to model and predict occupant's behaviour energy-related parameters affecting building energy performance are being sought for in this research. By accurately quantifying the impact of occupant's behaviour on energy performance, more reliable energy performance prediction can be achieved; hence a step forward towards supporting energy efficiency targets and related building policies and regulations.

1.3. Problem Statement

Predicting energy consumption in residential buildings with precision is crucial for climate change mitigation and sustainable development (Lim & Yun, 2017). However, a significant discrepancy between estimated and actual energy use persists, largely due to the current predictive models' simplistic assumptions about occupant behavior (Barthelmes, Becchio, et al., 2017). This gap not only undermines the economic and environmental efforts but also highlights the necessity for models that capture the nuanced dynamics of human interactions with their living spaces.

Studies have established the profound influence of occupant behavior on building energy performance. Yet, there is a clear shortfall in integrating the full spectrum of these behaviors into predictive models, which limits the understanding of their effects on energy consumption. The focus of existing research on a select few active behaviors, like lighting (Gunay et al., 2017) and thermostat adjustments (Gunay et al., 2018), which doesn't account for the interconnected nature of these actions within the broader ecosystem of residential energy use.

For instance, thermostat adjustments are not standalone actions but are part of a complex interplay involving internal heat gains from appliances, the physical presence of occupants, and external climatic conditions. Similarly, the decision to open windows for ventilation is a behavior that interacts with heating and cooling demands, personal comfort levels, and preferences for natural over mechanical ventilation.

This research responds to the need for an integrated approach that acknowledges the complexity and interdependence of occupant behaviors. It proposes an ensemble machine learning model that enhances the precision of energy consumption predictions by accounting for occupant behaviors. The model leverages the combined strengths of Lasso regression, Ridge regression, Random Forest, and Gradient Boosting to form a robust framework capable of capturing the intricate patterns of energy use.

1.4. Research Questions

The research questions below are devised to answer the hypothesis:

RQ1: What specific shortcomings exist within the current literature regarding the influence of occupant behavior on residential energy performance, and how can they be addressed?

RQ2: How to develop an occupancy behaviour-based ensemble machine learning model to improve energy consumption accuracy prediction of residential buildings using Lasso regression, Kridge regression, Random forrest, and Gradient boosting?

RQ3: How can the accuracy of the proposed occupant behavior-inclusive ensemble machine learning model be assessed?

RQ4: How to validate the reliability and real-world applicability of the proposed ensemble machine learning model for energy prediction?

1.5. Research Aim and Objectives

1.5.1. Research Aim

The aim of this research is to develop an ensemble machine learning model informed by occupant behavior to narrow the discrepancy between predicted and actual energy consumption in residential buildings. By comprehensively incorporating the various factors that influence occupants' behaviors into

energy prediction models, the research seeks to enhance the precision and reliability of these predictions.

The proposed ensemble model aims to offer energy modelers a comprehensive framework that enhances the precision of energy performance predictions by factoring in the multifaceted interactions of occupant behaviors. This framework is intended to guide informed design and operational choices that bolster energy efficiency within residential environments.

Hypothesis:

Developing an occupancy behaviour-based ensemble machine learning model will improve the accuracy of energy performance predictions in residential buildings compared to traditional simulation methods.

1.5.2. Research Objectives

The main objective of this research is:

"To Improve Energy Consumption Accuracy Prediction of Residential Buildings"

To achieve aim, the following objectives are formulated:

RO1: To identify most prominent occupants related parameters influencing the residential building energy performance, systematically review and evaluate the machine learning algorithms to determine the best fitting algorithm to be applied to predict the energy consumption.

RO2: To develop an ensemble machine learning predictive model for residential buildings energy consumption accuracy based on occupant's behavior-based inputs.

RO3: To evaluate the improvement of prediction accuracy of an ensemble machine learning model by comparing existing solo models.

RO4: To validate the applicability of the proposed ensemble predictive model with diverse evaluation metrics.

RO5: To validate the applicability of the proposed ensemble predictive model by applying in a real residential building and compare to simulation results.

1.6. Scope of the Research

In recent years, many researchers focus on predicting energy consumption in residential sector due to its importance in growth and economic development of the country. Because of imbalance use of energy in residential buildings due to seasonal and nonlinear behaviour of energy consumption, many countries suffer from energy crisis and some of the countries waste extra energy due to ineffective prediction of energy consumption. Hence there is a need to predict energy consumption accurately to cover the above-mentioned issues. There is a compelling need for improving accuracy of residential building energy predictions to be more occupant's behaviour oriented to provide more reliable energy predictions and hence minimize the performance gap. Improved prediction accuracy along with low error rate can significantly contribute to the economic development of the country. Resultantly, not only energy wastage significantly reduces but energy crisis can also significantly have covered.

The main contribution of this research is to propose an occupancy behaviour-based ensemble machine learning model to reduce the gap between actual and predicted energy consumption residential buildings by integrating all factors contributing to occupants' behaviours into building energy predictions. The inclusion of occupant's behaviour impact on energy consumption is the focus of this research as this serves in minimizing the energy performance gap and provide more reliability in predictions.

1.7. Significance of the Research

A significant amount of energy is wasted every year due to incorrect usage which can be reduced by using energy effectively. To reduce energy wastage intelligent solutions are required which can be overcome using accurate energy consumption prediction. Minimizing the energy performance

gap has been a challenge over the past decades for researchers and buildings energy modelers, achieving the anticipated performance and abiding to energy codes is indispensable.

To increase building energy utilization, a specific effective strategy must be put out. Building managers may make better decisions and more effective use of all sorts of equipment by using building energy consumption prediction. As a result, this approach is efficient and useful for reducing building energy use and raising energy utilization. Ineffective use of energy leads in a considerable annual loss of energy; hence waste may be decreased by doing so. There is also an imbalance in energy usage that can be the result of energy crisis as energy demands are increased with the development of smart systems in residential and industrial sectors. Hence there is a need to predict energy consumption using intelligent computing which is possible to predict with machine learning models. Due to the importance of energy consumption prediction, many researchers focused on this area very actively. Many researchers develop solo models to predict energy consumption and provide results with improved accuracy and low error rates but only a few researchers focused on ensemble models. Due to the best of knowledge, this area is still thirsty and needs attention to meet the desired accuracy with low error rates. The inclusion of occupant's behaviour impact on energy consumption is the focus of this research as this serves in minimizing the energy performance gap and provide more reliability in predictions. The proposed ensemble model will provide guidelines for energy modellers to improve the accuracy of energy predictions and understand all possible scenarios and outcomes of occupant behaviour in way informed design decisions can be deduced to promote energy efficiency in residential buildings.

1.8. Research Challenges and Limitations

- Sample Size Determination: The unpredictable and heterogeneous nature of human behavior complicates the process of choosing a representative sample size that reflects the diverse biological, physical, and social traits of occupants (Alaaeddine & Wu, 2017).

- Monitoring Frequency: The need for data collection frequency varies; activities like shade adjustment or window usage demand more frequent observation compared to longer-term behaviors like appliance usage (Fekri et al., 2021).
- Data Collection Period: While occupant behavior data is typically gathered over set periods, ranging from days to multiple years, this often necessitates generalization, despite the need for ongoing monitoring to capture the nuances of seasonal behavior changes (Candanedo et al., 2017).
- Ethical and Privacy Considerations: The collection of occupant behavior data is fraught with ethical and privacy issues, as well as organizational, legal, and practical challenges that limit the scope of data gathering (Delzendeh et al., 2017).
- Sensor Complexity and Cost: Deploying sensors for detailed behavior data collection can be prohibitively expensive and technically complex which is the reason for using existing datasets (Jiang et al., 2021).
- Survey Reliability: The reliance on surveys and questionnaires introduces the risk of inaccurate data due to the potential for occupants to misreport or incorrectly remember their actions (A. C. Menezes et al., 2012).
- Unacknowledged Constraints: The research must recognize and account for additional factors that influence occupant behavior, such as noise pollution affecting window usage, and the specific constraints of different building types. This tends to be a challenge due to limited data (Alaaeddine & Wu, 2017).
- In-depth Study of Occupant Characteristics: There is a gap in the extensive study and quantification of occupant characteristics as factors influencing energy consumption (Zou et al., 2018).

1.9. Research Gap

The research gaps investigated in this thesis are described below:

GAP 1: The review reveals a higher number of studies is required to gain confidence in the reliability and competency of occupant

behavioral patterns detection, and prediction of energy performance.

GAP 2: There is a scarcity of adaptive and reusable modelling approaches that can be used to predict the occupant's behaviour different buildings or different occupant's selection.

GAP 3: Most of the studies are focused on commercial and office settings which presents a gap in scrutinising other building types, this often relates to the unavailability of data, and lack of access in other building types or privacy concerns in such the case of residential buildings.

GAP 4: Lack of acknowledgement of additional constraints can impact occupant's behaviour such as noise pollution (windows opening) and building type related constraints is another finding from this research. Moreover, Occupant's characteristics (social, biological, etc.) are not studied extensively and quantified as contributing factors.

GAP 5: The different types of studies have been conducted on energy consumption prediction accuracy categories. However, complex ensembles including random forest, gradient boost, lasso, and ridge have not been ensembled with higher accuracy so far in the research literature. Researchers are showing increased interest in exploring ensemble techniques for achieving accurate energy consumption prediction results based on investigations that have revealed the accuracy improvement potential of this technique. It is also found that an ensemble technique produced better prediction results than a solo techniques.

1.10. Operational Definitions and Technical Terms

The methodology of this study is grounded in quantitative analysis. However, it does engage with a number of terms that possess expansive definitions within the scholarly discourse. These terms are presented in Table 1.2.

Table 1-2 Operational definitions

Terms	**Definition Description**
Energy Behavior	Occupants' activities that affects energy consumption of a building whether actively or passively.
Active Energy Behavior	The deliberate and intentional actions of occupants that affect the energy usage of a building, including the use of appliances, electricity, hot water, and the opening of windows, fall under the category of activities that impact energy consumption.
Passive Energy Behavior	The unintentional activities, particularly the production of metabolic heat, which can impact the energy consumption of a building.
Occupancy	The state of being present in/ or to occupy a space.
Energy Consumption	The use of energy simulation tools to predict building energy consumption by incorporating realistic inputs collected from primary data.
EnergyPlus	EnergyPlus is a building energy simulation software that is supported by the United States Department of Energy (DOE) and administered by the National Renewable Energy Laboratory (NREL). Engineers, designers, and researchers make

	use of EnergyPlus to forecast energy and water consumption, including heating, cooling, ventilation, lighting, and electricity usage, in buildings.
Lasso Regression	Lasso regression is a type of linear regression method that uses L1 regularization technique.
Ridge Regression	Ridge regression is a type of linear regression method that uses L2 regularization technique.
Random Forest	Random forest is a machine learning algorithm that uses an ensemble of decision trees to improve the accuracy of the prediction by reducing overfitting and increasing the robustness.
Gradient boosting	Gradient boosting is a machine learning algorithm that builds a sequence of weak learners, which are decision trees, to improve the accuracy of the prediction by reducing bias and variance.

1.11. Thesis Structure

Table 1.3 displays the correlation between the research questions and the chapters they are addressed in, as well as the alignment of the research objectives in the subsequent chapters.

Table 1-3 Mapping objectives and research questions

RQs	2	3	4	5	6
1	RQ1/RO1				
2			RQ2/RO2		
3					RQ3/RO3
4					RQ4/RO4/O5

There are six chapters in this thesis. The overview of the remainder chapter of thesis is shown in Figure. 1.9.

CHAPTER 1
INTRODUCTION

This chapter includes Introduction to research, research background, problem statement, research questions (RQ), research objectives (RO), scope, significance of research, Challenges and limitations, research gaps and operational definitions.

CHAPTER 2
LITERATURE REVIEW

This includes a comprehensive review on the prediction of energy consumption in buildings (tools and methods), the gap between the actual and predicted energy consumption in buildings and the impacts of occupants' passive and active behaviours on energy consumption in buildings are presented.

Besides this, the review of the determinants of energy consumption in residential buildings, occupant's behaviour and energy performance gap, review of present building energy performance simulation approaches, current approaches to modelling and predicting occupants' behaviours, Machine learning approaches, prediction of occupant behaviour by means of machine learning, gaps and findings.

CHAPTER 3
RESEARCH METHODOLOGY

This chapter contains a review of methods used to study the impacts of occupants' behaviours on building energy consumption, followed by, detailed description of

research method employed in this study including research philosophy, research approach, methodological choice, research strategy model development, data collection techniques and evaluation methods.

CHAPTER 4
MODEL ARCHITECTURE

Diving into the Algorithm selection, the model's description, modelling process in details.

CHAPTER 5
MODEL DEVELOPMENT & EVALUATION

In this chapter, the model is developed following steps detailed in the previous chapter. The results and evaluation take place.

CHAPTER 6
Model validation

This chapter shares the case study development. It includes further discussions on the results achieved through the ensemble machine learning model, validation of improvement in the prediction accuracy of energy consumption in residential buildings by incorporating occupants' realistic energy behaviours.

CHAPTER 7
Conclusion

This chapter contains the conclusion of the research linked with research objectives, in addition to future work.

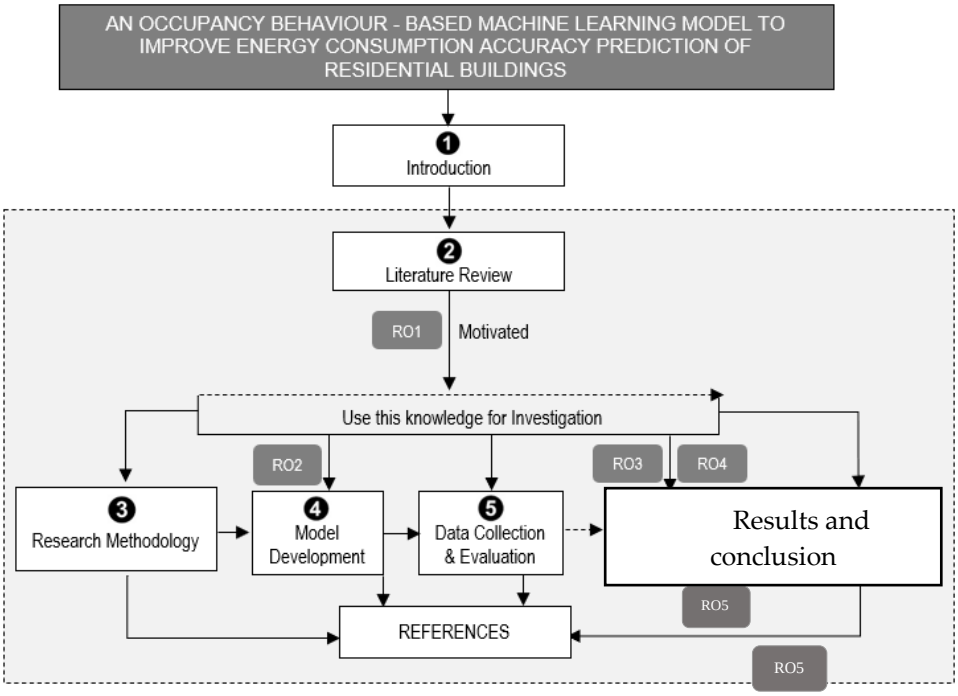


Figure 1-9 Research structure

AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE:
CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

Literature Review
Chapter 2

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

There is a significant emphasis on building energy performance and the prompting of energy efficiency measures amongst the scientific research community. Buildings are one of the primary energy consumers with an estimate of one-third of total energy resources (Paone & Bacher, 2018). The building energy use has been significantly rising over the last decade, which accounts for 40% of total energy use and the electrical use for the residential and commercial buildings accounting for almost 60% of the total electricity use Shabani and Zavalani (2017). Thus, the demand to reduce building energy consumption are high. Therefore, more energy efficient building design and operational solutions have been put in place Cao et al. (2016), which are analysed through energy performance simulation, modelling and prediction.

Energy performance prediction enables the exploration of different scenarios and investigating various solutions to reach optimised energy performance and utilise the energy in buildings in the most effective manner. Since the predicting the building energy performance has been established as a means to achieve energy conservation and explore alternative scenarios to promote effective building use the process of building energy performance prediction is always under study and development to become more reliable in reinforcing building energy optimisation and decision making (Huang et al., 2014). Building energy performance prediction depends heavily on the well-informed identification and understanding of all the parameters contributing to the building energy consumption and their integration in the prediction model. Those parameters include building physical and thermal characteristics, indoor/outdoor climatic and metrological conditions, building control systems

and services, indoor environmental quality requirements, and occupant related inputs (Krstić & Teni, 2017). The inclusion and reliable estimation of all these parameters, and the comprehension of their impact on building energy performance would result in more accurate energy prediction, which could support minimising the performance gap, and optimising the building energy performance.

The energy performance gap is disclosed as the gap between actual and predicted energy performance (De Wilde, 2014). One of the most influential factors contributing to the gap is said to be occupants behaviour and the energy estimates based on unrealistic, oversimplified and unreliable energy-behavioural assumptions (Buso et al., 2015; De Wilde, 2014). Providing accurate energy predictions implies great recognition and comprehension of occupant's behaviour in buildings and their impact on energy performance, this has proven to be challenging and complicated as occupants have stochastic and sophisticated behaviours (Yan & Malkawi, 2013). The realization of the impact of occupants behaviour on energy performance calls for providing alternative approaches to predict energy performance that can learn from the interaction between the occupant and the building (Alaaeddine & Wu, 2017). Numerous attempts to incorporate the impact of occupant's behavioural parameters affecting the building energy performance in simulation models have been explored to provide more accuracy and reliability (Jang & Kang, 2016; Wang & Ding, 2015). The prediction attempts testified that building energy performance calculation has a nonlinear dependency on a various number of external and internal variables with high-level dimensional data (Huang et al., 2014).

Machine learning approaches have been taken into consideration as prediction approaches for their ability to map nonlinear dependencies and deal with high dimensional data. Machine learning approaches can deliver otherwise challenging tasks, such as classification of behavioural patterns, and providing insights and prediction from complex data sets, which enables to answer critical questions about the impact occupant's behaviour on energy performance. Therefore, machine learning approaches to model and predict

occupant's behaviour energy-related parameters affecting building energy performance are being sought for in this paper to establish the current state of the art machine learning approach in predicting the impact of occupant's behaviour for energy analysis.

2.2 Occupants Behaviour and Building Energy Performance

Occupant behaviour impacts the building energy consumption; the way that occupants act, interact with building affect energy utilisation (Barthelmes, Becchio, et al., 2017). Occupant behaviour in buildings is reported as one of the most influential factors of the energy performance gap (Hong et al., 2017). Other influential parameters contributing to the performance gap are building physical characteristics and systems, load calculations, climatic and weather data which have been widely studied (Delzendeh et al., 2017). However, the inaccurate representation and quantification of occupant's energy-related behaviour in buildings remain the critical factors that require further investigation. Figure 2.1 provides an overview of the parameters affecting the building energy use, which undoubtedly complicate the energy usage prediction, while Figure 2.2 provides an overview on the parameters influencing occupant behaviour and in return the building energy use.

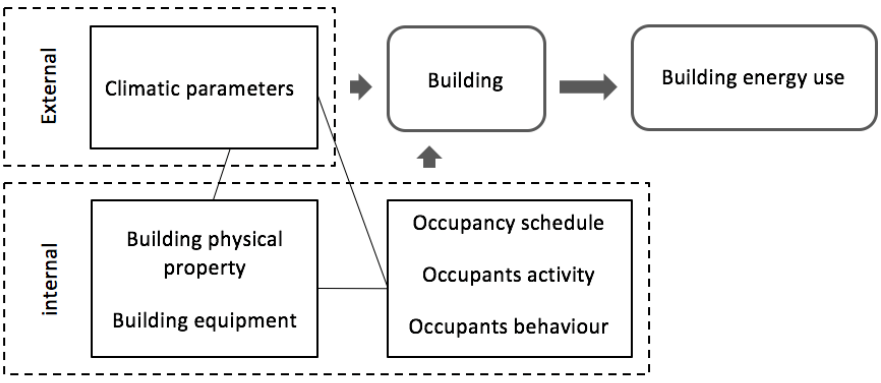


Figure 2-1 Parameters influencing building energy use

Reasons behind the difficulty of occupant behaviour prediction lie among the uncertainties accompanying human nature, the difficulty in identification of the drivers and needs affecting occupant's behaviour, the limited ability to quantify and estimate occupant's behaviour, and the lack of detail and abundance of data related to occupant's behaviour. Also, occupant

behaviours representation in the building performance simulation tools relies on assumptions, predefined schedule, oversimplified inputs, fixed settings and deterministic rules which conflicts with the stochastic and diverse nature of the actual behaviour (Yan et al., 2015).

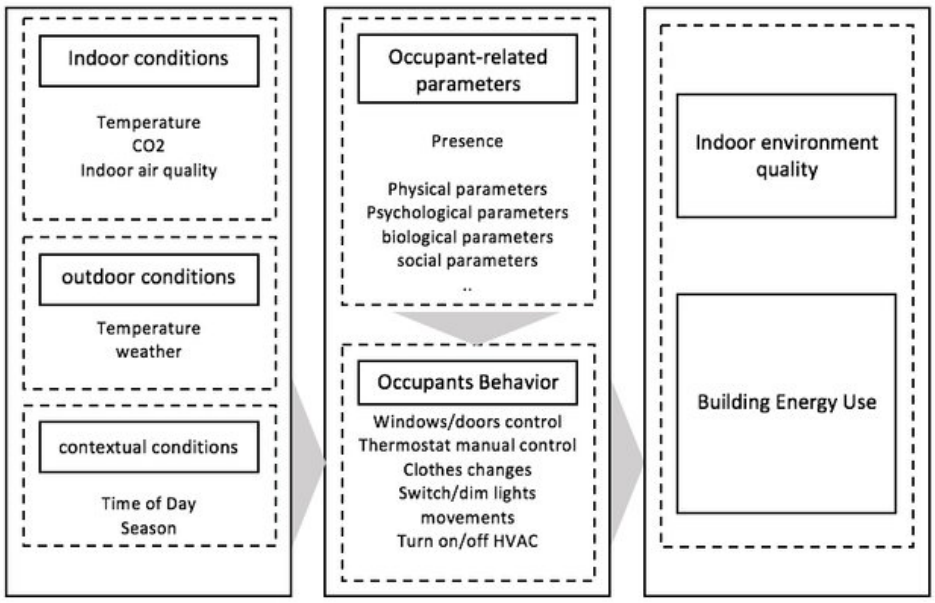


Figure 2-2 Effect of occupant behaviour on building energy performance

For reliable energy performance predictions, better representation of occupant behaviour is essential. Numerous studies explored the energy-related behavioural aspect of occupants and its impact on building energy performance. Reported variation in building energy use based on occupants behaviour is present in numerous studies (Fabi et al., 2013; A. C. K. d. Menezes et al., 2012). The factors leading to the variations in occupants behaviour are also analysed; such as occupants lifestyle (Barthelmes, Becchio, et al., 2017; Becchio et al., 2016), occupancy and household characteristics, occupants characteristics such as gender and age (Indraganti et al., 2015), drivers, attitudes, needs and values (D'Oca et al., 2014; Fabi et al., 2016; Hong et al., 2015; Huebner et al., 2015). Moreover, motivating drivers of occupants behaviors were explored in an attempt to provide standardized quantitative descriptions. For example, (Nicol & Humphreys, 2010) looked into Adaptive comfort theory. Adaptive comfort theory drivers of occupants' behavior were explored in an attempt to provide standardized quantitative descriptions. It relates that occupants can tolerate greater

fluctuation in acceptable temperature ranges when they are adapted to their environment. The study showed that allowing building occupants to interact with control systems leads to higher satisfaction and greater tolerance for fluctuations in acceptable temperature ranges. This can result in a reduction in energy consumption by up to 30% (Hong & Lin, 2013). Encouraging occupants to adopt comfort-adaptive energy-saving behaviors can be a cost-effective investment. This theory has been researched and tested to firmly verify that neutrality can be attained in indoor temperatures ranging almost 10 K above 21.6 degrees Celsuis, which is the productivity optimum. This all depends on the occupiers' prior thermal history (de Dear et al., 2013). In addition to that, looked into social practice theory. This theory has been utilized to inspect the impact of social behavior on energy use. It studied routine activities and behaviors, such as washing clothes, and taking a shower. The social practice theory highlighted that occupants can be shaped by their routines which is impacted by a combination of factors such as social norms, individual circumstances, technology, history. All of these which can affect how the occupant consume energy. Table 2.1 shows some theories that are prominent in understanding occupant behavior energy use in buildings (Higginson et al., 2015).

Table 2-1 Theories Influencing Occupant Behavior and Energy Use in Buildings

Theory	Description	Impact on Energy Use
Social Practice Theory (Higginson et al., 2015)	SPT considers the interplay between technology, social norms, and everyday life, and how routines evolve over time.	Provides insights into the influence of social norms on energy use behavior
Behavioral Comfort Theory (Paciuk, 1989)	This theory examines the relationship between the physical environment and occupant comfort, which can influence energy use behavior.	Understanding how occupant comfort preferences impact energy use
Theory of Planned Behavior (Conner & Armitage, 1998)	TPB posits that attitudes, subjective norms, and perceived behavioral control can influence behavior, including energy use.	Identifying factors that drive intentional energy use behavior
Diffusion of Innovation Theory (Kaminski, 2011)	This theory examines how new ideas or innovations spread through social networks, which can influence the adoption of energy-efficient technologies.	Identifying the factors that impact the adoption of new energy-efficient technologies
Maslow's Hierarchy of Needs Theory	This theory suggests that human needs are arranged in a hierarchy, with basic physiological and safety needs being the most important.	Understanding how basic human needs can impact energy use behavior

2.3 Impact of Occupants Behaviour

Occupants contribute to the energy use in building and affect the indoor environment through their presence and action in the buildings. Hong et al. (2016) defined occupant's energy-related behaviour as the interaction of occupants with the building, which involves the control of shades and blinds, adjusting the thermostats and HVAC systems, windows opening and closing, control of light, use of appliances, and occupant's movement between spaces shown in Figure 2.3. Moreover, energy-related behavioural adaptations include the adjustment of clothing, change in metabolic rate and consumption of drinks which affects occupant's perception of comfort and consequently influence their actions and energy performance.

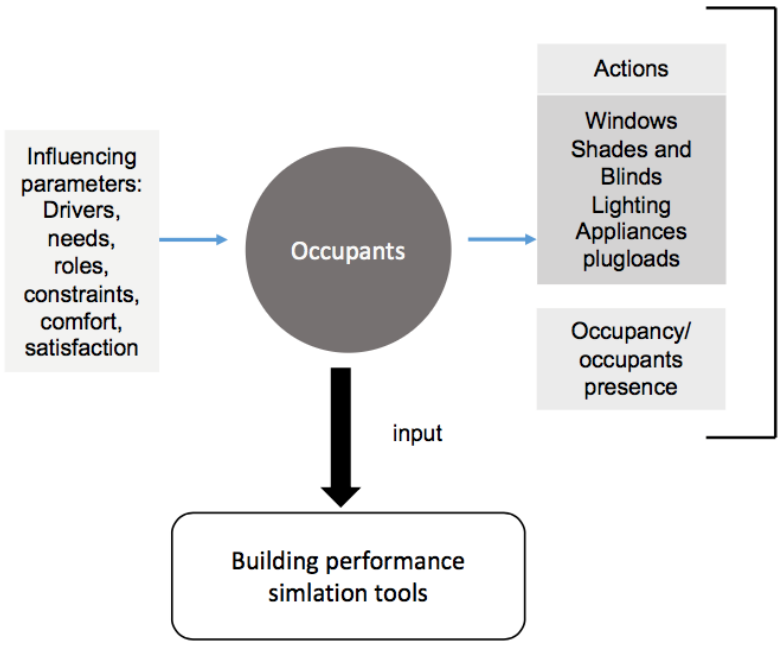


Figure 2-3 Occupant's behaviour

Delzendeh et al. (2017) categorised the effect of occupant's behaviour into two categories:

i. Passive effect - occupancy, heat and moisture gains from occupants
ii. Active effect - occupant's interaction with the building and its systems

The passive and active behaviours can be summarised as follows:

2.3.1 Window Opening and Closing

Windows provide occupants with a mean to control the visual and thermal comfort level by adjusting the window state (open/ajar/close). Windows operation and state is related to providing appropriate indoor air quality and thermal satisfaction as per the occupant's preference (Bruce-Konuah, 2014). Changes in ventilation rates are present when an occupant change the window state. The operation of windows has proven to have a significant impact on building energy use. As a result, this impacts on the overall building energy performance (Olu-Ajayi et al., 2022a).

2.3.2 Shade and Blind Operation

Shades and blinds control provide the occupant with a mean to adapt to their visual and thermal comfort level and satisfaction concerning privacy, daylight levels, glare, solar gains, as well as workplace illuminance. Shades and blinds could be controlled manually or mechanically depending on the devices in the building. The adjustment of shade or blind angle, position, and tilt, as well as the rate of interaction of the occupant with the shades and blinds, affect building energy use, lighting use, peak loads, and consequently building energy performance (Truong et al., 2021).

2.3.3 Lighting Control

Dimming and switching on/off lighting devices are based on occupants physical and visual comfort levels and satisfaction. The control of lighting could also be affected by occupants energy awareness levels, workplace policies and social factors (Somu et al., 2020).

2.3.4 Thermostat and HVAC Adjustment

The adjustment of thermostat set points and control of the HVAC system are the result of occupant's thermal comfort preferences. The accessibility to the control system depends on building types and size; for example, occupants have complete access and control to thermostat

adjustment in residential units while lacking control in multifunctional public spaces. The adjustment of thermostat set-points and HVAC control has a direct effect on building energy use (Perera et al., 2014).

2.3.5 Appliances Usage

The use of building appliances contributes to building energy use by influencing electricity consumption. The appliances also have an impact on heat gains, which in turn promotes other occupant's behaviour. The use of appliances is determined by the occupant's needs and type of activity. The prediction of plug loads contributes to building energy performance predictions (Hong et al., 2016).

2.3.6 Occupancy and Occupant's Movement (Passive)

Occupants' actions and interaction with building systems are dependent on the presence of occupants in the building in the first place. The occupancy of space, occupant's movement and density altogether impact building energy performance. Indoor air requirements, internal heat gains and energy consumption in a building are affected by the occupant's presence, schedules and density within building spaces (Dong et al., 2021).

2.4 Techniques for Predicting Energy Consumption

2.4.1 Simulation Techniques

The direct input or control approach establishes the semantics of occupant-related inputs, similar to other model inputs like building geometry, constructions, internal heat gains, and HVAC systems. In this method, users input and define temporal schedules for thermostat settings (cooling and heating temperature set points), occupants, lighting, plug loads, and the HVAC system. This approach necessitates that users compute schedules beforehand based on the correlations between environmental conditions and occupant actions in the occupant behavior models. Occupant behavior pre-calculation outputs rely on pre-defined rules, default values, or assumed environmental

conditions. Users might need to manually adjust the pre-calculation assumptions based on simulated results several times to ensure their accuracy.

This can be challenging, particularly when certain dynamic indoor parameters (e.g., air temperature) are used on both sides of the correlation function (e.g., switching on or off air conditioners when feeling hot or cold). In this approach, static set points (e.g., temperature set point) are commonly employed as an approximation to determine occupant actions and create schedules, which might reduce the accuracy of the occupant behavior models (Zendeh, 2019).

The second approach involves utilizing occupant behavior models, typically within a dedicated module. Although the built-in occupant behavior models approach offers a straightforward way to model specific the parameters in models, the limited availability of built-in models restricts its flexibility.

In the user function or custom code method, users can create functions or custom code within a building energy model input file to implement new building operation and supervisory controls or override existing or default ones. For instance, EnergyPlus features an energy management system, and DOE-2 has a user function feature that provides such functionality (Yan et al. 2015). This technique affords flexibility by permitting users to simulate a building energy model without recompiling the source code. It caters to both deterministic and stochastic occupant bevahior models, employing built-in or custom-developed stochastic mathematical functions (Delzendeh et al., 2017).

Co-simulation is a simulation approach that enables various components to be modeled by separate simulation tools running concurrently, exchanging information within a unified process (Wetter, 2011). Presently, the most sophisticated visual comfort and blind control models rely on image-based annual glare analysis from numerous perspectives within a scene. These models employ a combination of tools, such as RADIANCE, DIVA-for-RHINO, DAYSIM, Dialux, EVALGLARE, and other relevant software, to achieve a comprehensive understanding of the scene and its visual comfort aspects (Gunay et al., 2014). Also, instances can be found in computational fluid dynamics based studies on natural ventilation, which predict the performance of large-scale, naturally ventilated buildings with tools such as ANSYS-fluid, COMSOL, and Autodesk CFD (Wang et al., 2008). In conclusion, simulation

tools play a critical role in understanding and predicting occupant behavior and building performance. The availability of various approaches, such as direct input or control, built-in occupant behavior models, user functions or custom code, and co-simulation, provides researchers and energy modelers with options to tailor their analysis according to specific needs. However, It is crucial to consider the limitations and strengths of these approaches. As building design becomes more complex and occupant behavior patterns continue to evolve, advancements in simulation tools and the integration of occupant behavior models will be vital for creating energy-efficient, comfortable, and sustainable built environments.

2.4.2 Machine Learning Techniques

Machine learning is statistical based computational learning, which utilises the theory of statistics to provide mathematical models through making an inference from a sample dataset. In simple terms, it is inferring knowledge from data; and its learning is the execution of a computer program to optimise the parameters of a model using the training data or experience.

There are various types of machine learning techniques applied for various purposes; such as predictive types that make predictions on the future forecasting, and descriptive types which gain knowledge from data to identify risk factors (Kodratoff, 2014). Machine learning algorithms can be broken into supervised and unsupervised as shown in Figure 2.4. Supervised (predictive), infers a function from labelled training data, knowledge is present on input and output data (previous history on the input data, such as subject-related data and performance measures and performance results). Both input and output data are used to make predictions. In supervised learning, the algorithm is trained, and at the end of the process, the function that best describes the input data is selected. Training data containing the input/predictors is introduced, as well as the output data, and from this data, patterns and insights are presented. Supervised learning algorithms intend to model relationships and dependencies between the target prediction output and the input features such that predictions of the output values for new data based on those relationships learned from the previous data sets are available.

On the other hand, Unsupervised algorithms train with unlabelled data for pattern detection *and* descriptive modelling. These algorithms try to use techniques on the input data to mine for rules, detect patterns and summarise and group the data points which help in deriving meaningful insights and describe the data better to the users. The main types of unsupervised learning algorithms include *Clustering algorithms, association rule learning algorithms,* k-means for clustering problems, and Apriori algorithm for association rule learning problems.

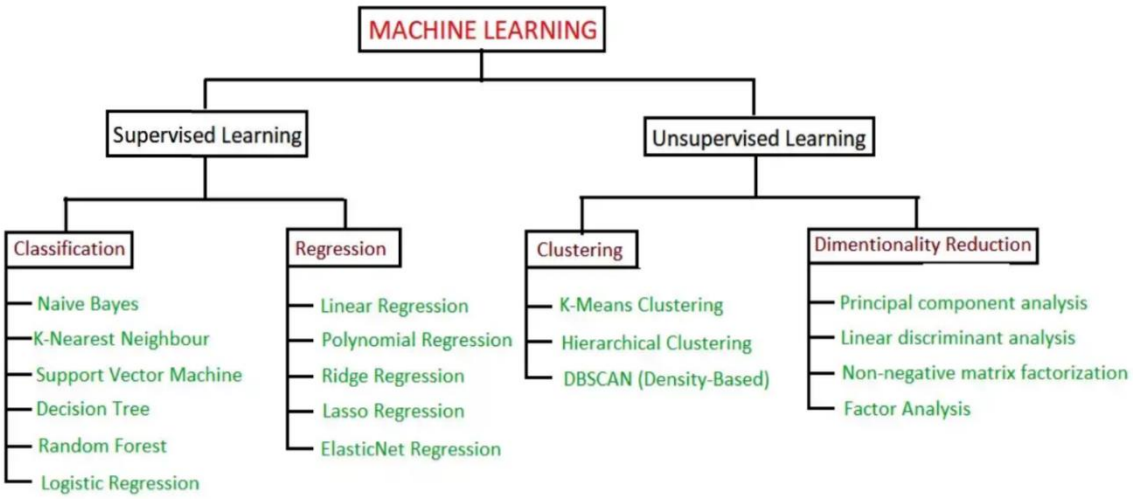


Figure 2-4 Overview of machine learning models

A brief description of common algorithms for machine learning algorithms and their applications are presented as follows.

2.4.1.1 Linear and Logistic Regressions

It is a statistical technique used in the fields of finance, investing, and other discipline that aims to establish the nature and strength of the relationship between a single dependent variable and several independent variables. Regression algorithms are the most common algorithms for general and simplified predictions. One of the common and early uses of regression applications is for load predictions such as electrical loads. Regressions are

built on historical data and aim at explaining the relationship between one dependent variable and a set of independent variables using linear combinations of the latter or estimating probabilities using underlying logistic functions (Tso & Yau, 2007). Regressions are used to find the impact of a variable based on historical data and predict future scenarios using new datasets. Linear regression is used for continuous targets, while logistic regression is for categorical targets. For energy predictions, some of the examples of the application of regression algorithm are to predict cooling loads Li and Huang (2013), in which multiple linear regression is used to relate predicted cooling load to multiple input variables including outdoor air temperature, solar horizontal radiation, room temperature set point etc. The use of multiple linear regression showed high-level accuracy and precision in their study when measuring prediction accuracy and precision (Mean bias error and coefficient of variance). Zhao et al. (2013) employed linear regression to predict the space occupancy schedule based on total energy consumption, which showed feasibility and applicability in prediction results.

2.4.1.2 Bayesian Networks

Bayesian networks are graphical models demonstrating probabilistic relationships among a set of random variables. Bayesian network is applied to model and explain a domain, support decision making under uncertainty, and find most probable configurations of variables. Bayesian Networks assign probability factors to outputs according to an analysis of a set of input parameters.

According to Darwiche (2009), Bayesian networks involves the following components:

i. The structure of the network defined as a directed acyclic graph, in which the random variables are presented by nodes, while dependencies among variables are represented by directed edges.
ii. Conditional probability distributions assigned for the variables.

Bayesian networks have been frequently applied in the real world for forecasting, diagnosis, automated visions in general. Also, Bayesian networks have been applied to predict occupants movements, cooling and heating loads, and overall energy consumption (Yan & Malkawi, 2013).

2.4.1.3 Decision Tree

Decision tree is considered one of the main algorithms for classification and prediction tools due to its hierarchical structure. A decision tree is considered a hierarchical model consisting of a set of decision rules that recursively arranges the input parameters into homogeneous zones. The decision tree can be a regression or classification tree. Its purpose is to provide a prediction by defining a set of decision rules based on the input parameters. Decision tree deals with the interaction between parameters and provides high efficiency with low computational effort (Singh et al., 2016). Decision tree allows the extraction of needed information from databases and has been widely employed in for business, predictions, and management.

2.4.1.4 Support Vector Machines

Support Vector Machine (SVM) is a machine learning algorithm that provides a sparse pattern of solutions and flexible control on the model complexity. It is commonly applied to map original input variables into high-dimensional feature space which introduces the non-linearity in the solution(Singh et al., 2016). SVM has shown high capability in dealing with classification problems in many fields, namely medical and bioinformatics. Also, it has been applied to predict building energy use in various case studies (Paudel et al., 2015; Solomon et al., 2011).

2.4.1.5 Artificial Neural Network

Artificial Neural Network (ANN) techniques are nonlinear statistical learning techniques resembling the biological neural configuration. ANN has the remarkable capability in modelling complex and nonlinear patterns. It has

been applied widely to predict energy use, and occupants' movements. A standard ANN architecture consists of input, output, and hidden layers. ANN is employed as a random function approximation tool that can capture complex relationships between inputs and outputs and model dynamic problems. As such, ANN provides ease of use in modelling problems that are difficult to explain (Wang & Srinivasan, 2017).

2.4.2.6 Ridge Regression

Ridge regression is a statistical technique used for examining data with multi-linearity. Ridge uses a penalty to define the magnitude of the coefficients, leading to their shrinkage toward zero. This, subsequently, leads to a more steady and interpretable model. Ridge has been widely used to predict occupants' behavior in buildings, since it can provide high correlations by handling multiple predictors. The technique provided promising results in improving the accuracy of energy use prediction models (Ding et al., 2021; Wang et al., 2020).

2.4.2.7 Lasso Regression

Similar to ridge regression, Lasso also deals with data with multicollinearity and leads to a more sparse model by handling regularities. Lasso can handle high-dimensional data with a large number of predictors, which makes it a solid algorithm when looking at the high number of predictors related to building energy performance (Deng et al., 2018).

2.4.2.8 Gradient Boosting

Gradient Boosting is an ensemble machine-learning algorithm. It combines multiple weak models to build a strong predictive model. Gradient boosting aims to minimize residual errors by the addition of decision trees. Since it can handle complex relationships between parameters as well as non-linearity, this algorithm has been used widely to predict occupants' behavior in

building with promising improvement in accuracy to energy use models (Deng et al., 2018; Guo et al., 2023; Wang et al., 2019).

.

2.4.2.9 Random Forest

Random Forest is a notable ensemble machine learning algorithm widely used for delivering strong predictive models. RF combines multiple decision trees. Those decision trees are built by the random selection of subsets of features and data points. Those decision trees are then combined to make predictions. Since RF handles complex responses and relationship as well as accounting for the interaction between predictors, it has been used to predict energy performance while accounting for the complexity in occupants' behavior (Ahmad et al., 2017; Azar et al., 2022; Deng et al., 2018; Wang et al., 2018).

2.5 Literature on Occupant's Active Behaviour

The literature search identifies the four principal occupant's actions based on the number of studies shown in Figure 2.5. The four actions represent occupants active behaviour and have been studied extensively as these actions, and their combinations aim to improve or maintain the occupant's indoor environmental quality and comfort, consequently affecting the energy use in buildings (Fabi et al., 2011). The attempts to predict each of these actions through machine learning categories are studied and explored separately. Thus, this review and discussion are divided into four parts.

i. Window opening and closing
ii. HVAC control and thermostats adjustment
iii. Appliances use
iv. Shades, blinds and lighting control

A review of window control modelling attempt employing advanced statistical and machine learning approaches is presented in Table 2.2.

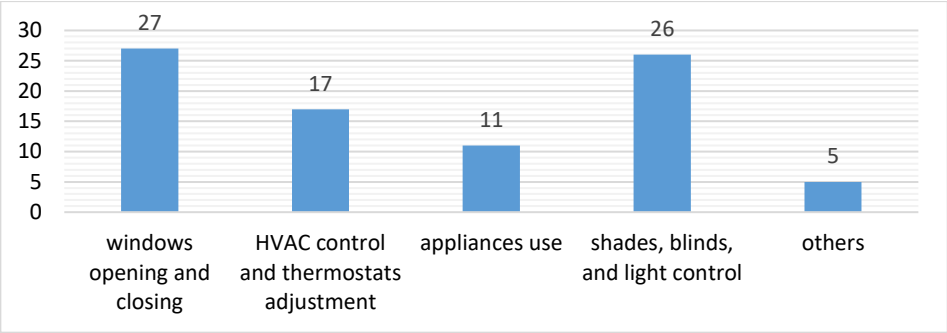


Figure 2-5 Number of studies for occupants' behaviour

2.5.1 Window Opening and Closing

The control of windows is a part of occupants behaviour that needs to be taken in consideration when simulating building energy performance, this way more reliable presentation of the occupant's response to the indoor environment quality, thermal comfort levels, and ventilation preferences are reflected (Fabi et al., 2012)[S1]. Having established that the control of windows for providing ventilation in the building is a parameter that needs to be considered, this parameter depends on the state of window decided by the building occupants which in turn is triggered by indoor and/or outdoor temperature, the activeness or passiveness of the occupant, amongst other factors. There have been several attempts to model and predict the occupant's control of windows and its impact on energy consumption (D'Oca et al., 2014; Wei et al., 2015)[S2,S3]. D'Oca et al. (2014)[S2], proposed a probabilistic modelling approach utilising multivariate logistic regression to combine the energy model probabilistic user profile for window and thermostat control was developed; significant findings of this study reiterated the lack of reliability of the oversimplified occupant profiles in the conventional energy simulation tools. These models aim to provide more adaptive window control schedule and more receptiveness to the environmental conditions demonstrated by actual occupants in buildings. Some of these models takes into consideration the window controlling behaviour of occupants based on the indoor/outdoor air conditions which neglects the individuality and stochastic nature of occupant's behaviour which is an outcome of behavioural drivers, needs and triggers based on a combination of physical, psychological, social, and comfort-related and preferential aspects of the occupants. Other models provide over-

representation of the window state being close; hence leading to biased results, or are tuned to specific occupants; hence introducing additional or different occupant to the model proves to be ineffective (Markovic et al., 2018)[S4]. The occupant control of windows and its impact on occupant's behaviour should reflect the stochastic nature and complexity of occupant's behaviour and cannot rely on deterministic approaches of modelling. Therefore, more modelling attempts of windows control are being researched and explored using machine learning and stochastic models. From the study of Table 2.1, publications involving windows opening prediction by machine learning has a steady flow over the years. However, it peaked in the year 2021 due to the rise of interest and awareness to the ability of machine learning to provide better predictions. It is also noticeable that office building and residential buildings have the highest share of the studies when predicting windows opening behaviour (43% each). However, only a small amount of studies involved educational buildings (2%), and another type of buildings such as hospitals and laboratories (2%) shown in Figure 2.6. This might be due to the limitation in monitoring and data availability for those buildings, and the restricted access of window control by occupants. However, a further study involving a wide range of buildings is needed to cover the prediction of windows control in various scenarios.

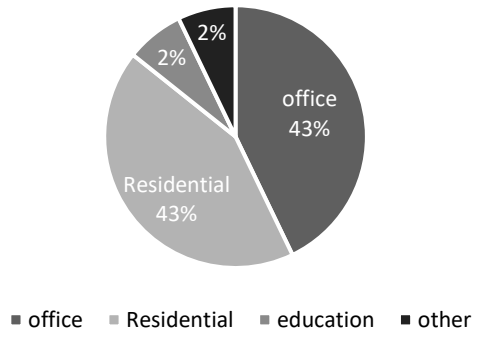


Figure 2-6 Types of buildings

Shi et al. (2018)[S5] studied windows opening/closing behaviour and their various influencing factors. Logistic regression models in variable seasons identified as cooling/heating/ transitional seasons are utilised for predicting the state of the window (open. Closed, ajar). The study realised that occupants respond to the indoor air condition and relative humidity, which affects their

window control in all seasons. The logistic regression models are validated, and the results show a promising accuracy level higher than 70%. Reported limitation of this study is the small size of sample data. Other studies (Kim et al., 2017; Li et al., 2015; Valentina et al., 2012; Yao & Zhao, 2017)[S6,S7,S8,S9] utilised logistic regression to provide predictive model to provide better understanding and prediction of the window control behaviour of occupants. The number of buildings, the nature of the study (field study, monitoring campaign, survey, questionnaires), and consequently the quality and availability of the data input deferred in each study. The results indicated that various variables impact the window opening/closing behaviour, and the models perform better when compared to the predefined and oversimplified schedules of window control behaviour provided in the conventional approaches. Li et al. (2015)[S7] revealed that bias is introduced when the outdoor temperature is not in-between 15 °C and 30 °C. Kim et al. (2017)[S6] deduced that temperature of 25 °C is the favourable temperature for window opening to provide natural ventilation, and personal and demographic characteristics have a significant impact on occupants use of mechanical ventilation. Most of these models did not include values influencing the window state that was to be predicted and rather relied on the environmental conditions as per the data available.

Dutton and Shao (2010)[S10] used a co-simulation approach with EnergyPlus to provide a more reliable simulation of window control behaviour; Logistic regression algorithms were utilised to develop the probabilistic model for a school. The aim was to determine the probability of proportion of window open, and it relied on the post-occupancy study and concurrent measurement of window state along with environmental conditions and energy use data. The behavioural predictions tended to provide more accuracy than the model taking into consideration the temperature set points. Calì et al. (2016)[S11] determined that there is a relationship between window opening and the concentration of carbon dioxide and the time of day, and between window closing and time of day and outdoor temperature. Their findings concluded that logistic regression provides a robust analysis method to study the occupants' window control. Although findings show that the regression models provided

more accuracy when compared to the standard simulation methods for window control, the models lack the depth of identification of behavioural patterns based on occupant related parameters and the underlying factors leading to the change of window state, the focus is more on the state of the window and its relationship to environmental conditions. This could be due to the limitations of data provided, the time and expenses for model training, and the complicated modelling approach.

Markovic et al. (2017)[S4] studied the performance of several classification algorithms for detecting occupant's behaviour in terms of windows control; the used algorithms are SVM, Random Forests, and their combination with Dynamic Bayesian Network DBN. The results show that Random Forest approaches had better performance accuracy when predicting the window state. Although all these approaches provided better accuracy than conventional prediction methods using the available simulation tools, limitations lie in the inability of learning every individual occupant's behaviour without an intensive training which requires more time and computational efficiency. Markovic also applied deep learning methods to predict windows opening (Markovic et al., 2018). The developed model was a multi-layered neural network which resulted in high-performance results and improved accuracy while maintaining low complexity levels. The limitations were in the sample size in the study and its lack of inclusion of different climate zones and various range of occupants. Kim et al. (2018)[S12] proposed a modelling approach for personal comfort employing machine learning algorithms (Classification tree, Gaussian process classification, gradient boosting method, SVM, random forest and logistic regression). The model aimed to predict individual thermal preferences. The logistic regression, random forest, and SVM have proven to provide a higher level of accuracy. These comfort models could provide a better understanding of occupants needs and hence understanding to their control of building systems such as windows control. The limitations lied in the size of the data sets, computational challenges and model complexities. Barthelmes, Heo, et al. (2017)[S13] attempted to capture the complicated relationship between window opening and various contributing factors using a Bayesian network model. The model showed promising

predictive accuracy for predicting occupants window control behaviour. (Stazi et al., 2017)[S14] assessed the influence of recorded parameters on windows status. The results revealed that daily routine and habits highly influence students' behaviours. (Jones et al., 2017)[S15] understanding the probability of opening and closing windows based on indoor and outdoor environment factors and according to the time of the day and season. The results concluded that indoor and outdoor relative humidity affected occupants' window operation behaviour.

Table 2.2 Studies of Windows control modelling

Study	Ref.	Purpose	Model	ML method	Building Type	Input Data	Output Data	Results
S1	(Fabi et al., 2013)	The procedure is applied at models of occupants' interactions with windows (opening and closing behaviour).	a probabilistic approach for modelling	Linear regression	20 simulations of the same model	indoor environmental parameters and external climate conditions and the behaviour of the building occupants (window opening, thermostatic radiator valves' set-point, occupancy sensors, etc.),	probability distributions of energy consumption and indoor environmental quality depending on user behaviour.	large variations range between behaviour patterns in the groups with natural ventilation and mechanical ventilation
S2	(D'Oca et al., 2014)	Combining probabilistic user profiles for both window opening and thermostat set-point adjustments into one building energy model	Probabilistic modelling	Multivariant logistic regression	fifteen naturally ventilated dwellings located 10 to 25 km from Copenhagen	field monitoring campaign of indoor and outdoor climate conditions and occupants control actions	Probability of control actions of window opening and thermostat set point	Major findings of this research demonstrated the weakness of standardized occupant behaviour profile in energy simulation tools
S3	Wei *et al.*, 2015	studying human adaptive behaviour in non-airconditioned buildings	Human adaptive models (adaptive preference model, occupancy scheduling model and behaviour determining model)		four non-air-conditioned buildings	Occupants' characteristics (gender, occupation, origins, location within building floors, distance to windows), preferences on adaptive behaviour,	Building performance simulation	valuable trends and potential influencing factors have been identified
S4	(Markovic et al., 2017)	Analyzing the performance of several classification algorithms for detecting occupants' interactions with windows	Several classification models	support vector machines, random forests, and their combination with dynamic Bayesian networks	Office building in Frankfurt Germany, over two years data.	Data include indoor climate and outdoor climate features, as well as occupants' presence and actions (occupancy. Occupancy status, time.)	Frequency and time duration of the window opening actions	random forests outperformed all alternative approaches for identifying the window status in office buildings.

S5	(Shi et al., 2018)	Studying window opening/closing behaviours and their different influencing factors	Stochastic models	Logistic regressions	two naturally ventilated hospital wards in Nanjing, China.	One-year field measurement, the effects of air quality and the climatic parameters on window opening/closing behaviours.	Prediction of the window opening/closing state	Model promised adaptable with results of accuracy bigger than 70%.
S6	(Kim et al., 2017)	better understanding residential adaptive thermal comfort behaviours	Predictive models	logistic regression models	42 households in two neighbouring Australian cities	Questionnaire surveys and instrumental monitoring indoor- and outdoor-climatic data Weekly Online comfort questionnaire, participants' demographic and personal information	predicting occupant adaptive behaviours based on different variables	The analysis indicated that an outdoor temperature of 25 °C was the most favourable condition, maximising the use of natural ventilation. The paper pointed out personal and demographic characteristics can have a significant impact on the householder's decision to use their air-conditioning system.
S7	(Li et al., 2015)	Investigating window-opening behaviour during the transition seasons when air-conditioner is inoperative	Monte Carlo simulation method	Logistic regression mode	A five-storey office building in Chongqing	Window opening status, indoor outdoor air conditions, climatic data	probability of window opening	When the outdoor temperature is beyond the range of between 15 °C and 30 °C, bias exists.
S8	(Valentina et al., 2012)	estimating the effect of the control on windows by different user behaviour patterns measurements of indoor climate and outdoor environmental parameters and window "opening and closing" actions	probabilistic models of inhabitants' window "opening and closing"	Logistic and linear regression	15 dwellings from January to August 2008 in Denmark.	Indoor environment parameters, Outdoor environment parameters,Window state (open/closed)	the probability of opening the window for four user profile, the degree of opening was then predicted.	the predefined schedule for the window control underestimates the opening and closing events compared to the probabilistic models
S9	(Yao & Zhao, 2017)	Determining factors influencing residential occupants' window opening behaviour and analyzing the	A stochastic model of occupants' window opening behaviour	multi-variate linear logistic regression	9 naturally ventilated residences in	survey the occupants with questionnaire to ascertain the factors that may drive them opening or closing the windows.	the "success" probability of window opening	The results indicate that influence of the identified variables on window opening behaviour was significant.

		relationship between probability of window opening and individual studied explanatory variables.			Bei-jing during spring	The occupied periods, monitoring the windows' status and environment factors.		
S10	(Dutton & Shao, 2010)	Simulating with energy plus to provide more realistic picture of window opening behaviour in the model	probabilistic model	Logistic regression	A naturally ventilated elementary school in the UK.	Post occupancy study; Concurrent measurement of window open state, CO2 concentration, temperature, and exterior environmental conditions, classroom daily occupancy levels and monthly building energy usage.	Probability of the proportion of windows open	Predictions of both CO2 concentration and building energy performance, using the occupant behaviour model, were shown to give more accurate predictions than a model based on temperature setpoints.
S11	(Calì et al., 2016)	Identification of drivers for window control	statistical method	Logistic regression analysis	three refurbished buildings	Air temperature, Relative humidity, CO2 concentration, Volatile organic compounds, Light on the ceiling, Infrared/visible light ratio, Window opening sensors (open/closed).	Drivers for opening closing windows	Logistic regression was confirmed to be a strong and logistic robust analysis methodology for investigating the drivers for occupants to interact with the built environment. The most common drivers for opening action are time of the day and CO_2 concentration. The most common drivers for closing action are outdoor temperature and time of the day.
S12	(Kim et al., 2018)	Predicting individuals' thermal comfort responses instead of the average	personal comfort model	Classification Tree, Gaussian Process Classification, Gradient Boosting Method, SVM,	Office building in California	field study examining the behaviour and thermal comfort perceptions of 38 occupants,	individuals' thermal preference	RF, SVM, logistic regression showed higher accuracy. The paper deduced that personal comfort models could

		response of a large population		Random Forest, Logistic regression.		PCS chair data, survey data, and environmental data		provide more accurate representations of occupants' comfort needs and desires.
S13	(Barthelmes, Heo, et al., 2017)	Capturing underlying complicated relationships between windows opening and various influencing factors	Stochastic modelling	BN model	residential apartment located in Copenhagen, Denmark	Outdoor and indoor variables and window status and time of the day	Windows opening behaviour	the validation measures confirmed the high predictive power of the model and its successful application for modelling window control behaviour.
S14	(Stazi et al., 2017)	Assessing the influence of recorded parameters on windows status	Adaptive behavioural models	Linear and logistic regression models	a high school in Italy	Outdoor and indoor conditions, users' adaptive actions And interaction with windows monitoring	probability of window opening/closing	Paper deduced that daily routine and habits highly influence students' behaviours
S15	(Jones et al., 2017)	Understanding the probability of opening and closing windows based on indoor and outdoor environment factors nd according to the time of the day and season	stochastic models	multivariate logistic regression	ten UK dwellings over a year	the physical environmental and contextual variables	Probability of the main bedroom window will be opened or closed in the next 10 min.	The paper concludes that indoor and outdoor relative humidity affected occupants' window operation behaviour.

2.5.2 HVAC Control and Thermostat Adjustment

The control of Heating, Ventilation, and Air Conditioning (HVAC) and thermostats adjustments is another parameter that needs to be considered when studying the impact of occupant's behaviour on building energy performance. This parameter is affected by numerous variables such as occupancy and density, clothing factor, thermal comfort preferences, indoor and/or outdoor temperature, the activeness or passiveness of the occupant, amongst other factors. The ASHRAE handbook Heating et al. (2000) takes into consideration the occupants control of thermostats and HVAC uses a defined temperature and humidity range as a thermal comfort preference for occupants. More consideration of the stochastic nature and complexity of occupant's behaviour need to be addressed to provide more accuracy. A review of HVAC and thermostat adjustments modelling approaches by employing advanced statistical and machine learning approaches is presented in Table 2.3.

From the study of Table 2.3, publications involving thermostats adjustment and HVAC control prediction by machine learning have noticed an increase in the year 2017 and 2018 due to the awareness to the capability of machine Learning approaches to provide better predictions. Limitations in the diversification of case studies are noticed as per this review. The case studies only include residential and office buildings with residential buildings having the highest share of 63% while office buildings are having a share of 37% shown in Figure 2.7.

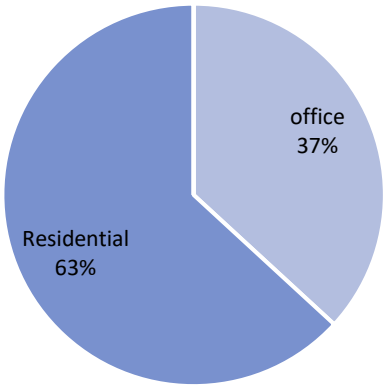


Figure 2-7 Building types

A wide range of Machine Learning algorithms is employed for predicting the thermostat and HVAC control by building occupants shown in Figure 2.8.

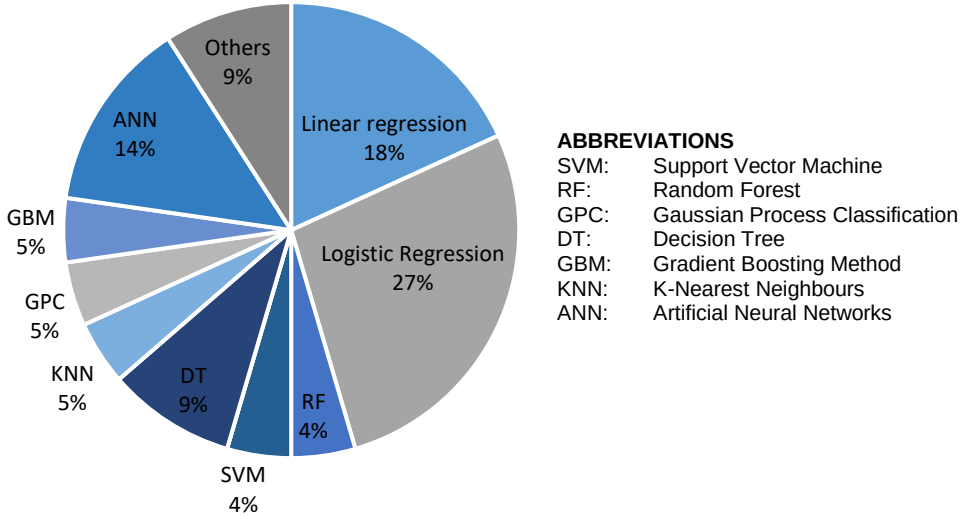


Figure 2-8 Machine learning algorithms for HVAC and thermostats adjustment

Regression approaches also represent the highest portion of models studies (45%). This relates to the lower level of complexity of regression models when compared to more advanced machine learning models. Gunay et al. (2018)[S16] studied the occupant's interaction with thermostats in private office spaces, by developing a univariate thermostat use model employing univariate logistic regression. The iterative learning process showed a 2°C–3°C reduction in the set-points during the heating season and a 2°C–3°C increase in the set-points during the cooling season concerning the default 22°C set-point. Santin (2011)[S17] attempted to identify the occupant behavioural patterns associated with the energy spent on heating and to determine the household and building characteristics that could contribute to the development of energy-User Profiles in residential buildings. The study demonstrated relationships between occupant behaviour and household characteristics. However, limitation lied in difficulty to demonstrate relationships between energy use and behavioural patterns and household groups. Also, a regression approach employed for a residential setting Tanimoto et al. (2008)[S18] proved that the deterministic approach of occupant schedule provides overestimated assumption when compared to the used approach to estimate cooling demands. Jain et al. (2016)[S19] applied a

statistical model to estimate the air-conditioning state using Linear regression. The model reached accuracies as high as 97% for the prediction of the air conditioning energy use in 7 houses in Delhi. Andersen et al. (2011)[S20] modelled heating set points preferences by employing linear regression approach based on simultaneous measurement of the set- point of thermostatic radiator valves in 13 houses in Denmark. Their study aimed to increase the reliability of energy simulations. Major findings included that the occupant's behaviour was governed by varied habits and the environmental variables are the important contributors when it comes to set point adjustments.

Advanced machine learning algorithms had a share of 55% of the reviewed studies. In one study, an artificial neural network model is developed to establish the relationship air temperature and relative humidity and occupants thermal-related behaviour (Deng & Chen, 2018)[S21]. Another study related to thermal related preferences used a number of machine learning algorithms (classification tree, Gaussian process classification, gradient boosting method, SVM, random forest and logistic regression) to develop a predictive personal comfort model) (Kim et al., 2018)[S22]. In addition to that, a neural network model is developed to study the thermal control strategies to create better thermal conditions and maintain occupants comfort (Moon & Kim, 2010)[S23]. These studies proved to have reliable accuracy and establish the link between the indoor environment and its effect on occupant behaviour and vice versa. Koehler et al. (2013)[S24] developed a hybrid model by employing the machine learning and preheat algorithms to match heating controls to occupants' preferences and routines. This study offered a chance to improve automated control systems and a mixed-initiative system for controlling occupant's thermostats. Kruusimägi et al. (2018)[S25] developed a heating control system that delivers thermal comfort and energy efficiency and evaluating its fitness for purpose in real-life contexts through learning users' room-specific presence profiles and thermal preferences. Simulation results confirmed that the algorithm functions as intended and that it can reduce energy need by a factor of seven compared with EnergyStar recommended settings for programmable thermostats.

Nägele et al. (2017)[S26] evaluated the performance of each heating control approach in terms of energy consumption and comfort for occupants using ANN-based thermal control models for residential buildings algorithm. The results revealed that automated setpoint variation; heating control approaches bear the potential to significantly increase energy efficiency in the residential sector. Ghahramani et al. (2017)[S27] simulated different control policies using EnergyPlus. The proposed algorithm resulted in 31.17% energy savings compared to the baseline operations (22.5 °C and 3 K). The algorithm has a superior performance in all climate zones for the goodness of measure. Lim and Yun (2017)[S28] investigated the implications of adaptive comfort control. Simulation with EnergyPlus Runtime Language (ERL) for modelling occupant behaviour. A dynamic thermostat control based on an adaptive comfort model is an effective method to reduce cooling energy consumption under future climate change.

Overall, the studies showed that machine learning approaches provide more reliability in predicting thermostats adjustment and HVAC control based on identifying occupant's thermal preferences. Limitations in these studies lie in consideration of occupant's traits and their impact on thermostats adjustment.

Table 2.3 Studies of HVAC control and thermostat adjustment

Study	Reference	Purpose	Procedure	ML Method	Building type	Input data	Output data	Results
S16	(Gunay et al., 2018)	Studying occupants interact with their thermostats	univariate thermostats use models	univariate logistic regression	private office spaces	occupancy, temperature, and relative humidity data	indoor temperature as the predictor variable	the iterative learning process resulted in a 2°C–3°C reduction in the set-points during the heating season and a 2°C–3°C increase in the set-points during the cooling season with respect to the default 22°C set-point in both seasons.
S17	(Santin, 2011)	To determine Behavioural Patterns associated with the energy spent on heating and to identify household and building characteristics that could contribute to the development of energy-User Profiles	Statistical analysis		Residential buildings	Household survey concerned with detailed data on occupant behaviour and paired wirh data on building characteristics	Behavioural patterns and user profiles	this study established clear relationships between occupant behaviour and household characteristics. However, it seems difficult to establish relationships between energy consumption and Behavioural Patterns and household groups.
S18	(Tanimoto et al., 2008)	estimating the cooling demand in residential context	Stochastic model	Linear regression	Residential setting	15 min activities of occupants, based on published data on occupant behaviour.	probabilistic variations in occupant behaviour, likelihood of switching air-conditioning on or off	the conventional procedure based on determinant calculation and a daily constant occupants' schedule results in incredibly huge overestimates compared with those produced by the novel method
S19	(Jain et al., 2016)	estimating Airconditioning state	Statistical model	Linear regression	7 homes in Delhi	2200 hours of usage data from the different ACs, room types, and thermostat temperatures	AC energy consumption prior to usage and estimate energy consumption post-usage.	The model achieved an average accuracy of 85.3% and 83.7% with the best accuracy of 97.0% and 93.3% for the estimation and prediction of AC energy consumption respectively, across all homes.

S20	(Andersen et al., 2011)	To Increase reliability of energy simulation by including occupants heating setpoints preferences	Statistical model	Linear regression	15 dwellings in Denmark	Simultaneous measurement of the set- point of thermostatic radiator valves, and indoor and outdoor environment characteristics	Variations in individual behaviour patterns	the behaviour was governed by different but distinct habits in the 13 dwellings. There are Correlations between environmental variables and set point on the thermostatic radiator valves
S21	(Deng & Chen, 2018)	to determine the relationship between air temperature and relative humidity, and occupants' thermal sensations and behaviour.	ANN model	Artificial neural network	Offices and apartments/houses	data on the thermal environment, thermal sensations, and occupants' behaviour; air temperatures, relative humidity, clothing levels, thermal sensations, thermostat setpoints, and room occupancy	predicting thermal comfort	The behaviour of occupants could be a significant parameter for evaluating indoor environments in buildings.
S22	(Kim et al., 2018)	predicting individuals' thermal preference	Personal comfort model	6 machine learning algorithms Classification Tree, Gaussian Process Classification, Gradient Boosting Method, SVM, Random Forest, Logistic regression.	Office building	field data including Personal comfort system control behaviour, environmental conditions and mechanical system settings	individuals' thermal preference	personal comfort models improved predictive accuracy compared to conventional models
S23	(Moon & Kim, 2010)	to develop residential thermal control strategies for creating more comfortable thermal conditions.	ANN models	Ann algorithms	typical two-story single-family home in the U.S	Setpoints, weather data, simulation results	Predict air temperature profile and energy consumption.	ANN-based predictive and adaptive control strategies created more comfortable thermal conditions

S25	(Kruusimägi et al., 2018)	Developing a heating control system that delivers thermal comfort and energy efficiency and evaluating its fitness for purpose in real-life contexts through learning users' room-specific presence profiles and thermal preferences	Heating control model	spatiotemporal heating control algorithm (unspecified)	three domestic homes	temperature set-point, occupants' thermal sensation feedback, occupant-dependant departure schedules	prediction of occupants' presence and preferred set-point temperature	simulation results confirmed that the algorithm functions as intended and that it can reduce energy need by a factor of seven compared with EnergyStar recommended settings for programmable thermostats.
S26	(Nägele et al., 2017)	Evaluation of the performance of each heating control approach in terms of energy consumption and comfort for occupants	Statistical model	ANN-based thermal control models for residential buildings algorithm	households in Southern Germany were collected over a 14-month period	data on in-room temperature, heating behaviour and occupancy patterns	Performance of heating controls	Intelligently. automated setpoint variation; heating control approaches bear the potential to significantly increase energy efficiency in the residential sector.
S27	(Ghahramani et al., 2017)	Optimizing energy use	Simulating different control policies using EnergyPlus	metaheuristic (k-nearest neighbour stochastic hill climbing), machine learning (regression decision tree), and self-tuning (recursive brute-force search) components	small office building	real-time data, stored in building automation systems (e.g., gas/electricity consumption, weather, and occupancy)	Optimal operation setting	The proposed algorithm resulted in 31.17% energy savings compared to the baseline operations (22.5 °C and 3 K). The algorithm has a superior performance in all climate zones for the goodness of measure
S28	(Lim & Yun, 2017)	investigating the implications of adaptive comfort control.	Simulation with EnergyPlus Runtime Language (Erl) for modelling occupant behaviour.		Office building in Seoul, Korea	Climatic, weather data, energy consumption, cooling loads.	quantification of the effects of adaptive comfort control on current and future cooling energy consumption in the context of climate change	a dynamic thermostat control based on an adaptive comfort model is an effective method to reduce cooling energy consumption under future climate change

2.5.3 Appliances Use

The use of appliances is an important factor when considering occupants behaviour and their impact on energy performance as appliances represent a noticeable percentage of the overall energy consumption (between 20 and 30%) (Cetin et al., 2014; Kavousian et al., 2015). The studies for appliances and plug controls represented the least amount of publications in this review. As for the building types, the limitations in the diversification of case studies are also present in the case of appliances use prediction with a percentage of 67% for residential buildings and 33% for office buildings shown in Figure 2.9.

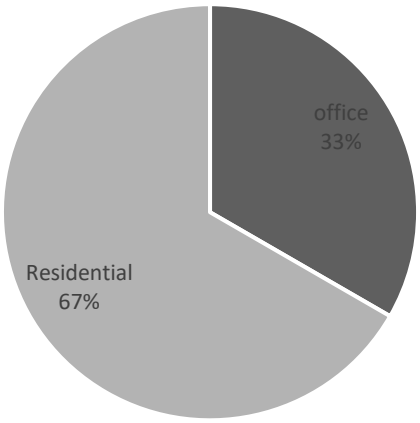


Figure 2-9 Building types

There is a wide range of machine learning algorithms used for the prediction of appliances use with varied percentages presented in Figure 2.10.

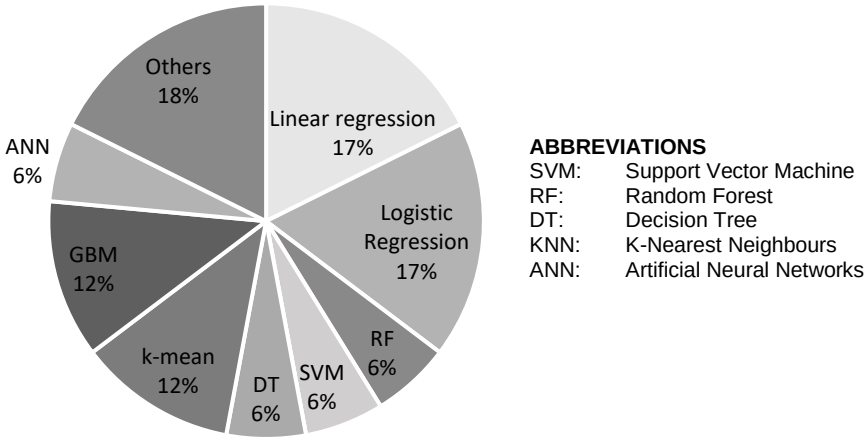


Figure 2-10 Machine learning algorithms for plug loads/appliances use

A review of appliances used approaches by employing advanced statistical and machine learning techniques is presented in Table 2.4. Wang et al. (2018)[S29] developed models using several machine learning algorithms to predict building electricity usage with part synesthetic case study office and an existing office building in Iowa. The algorithms employed included multiple linear regressions, adaptive linear filter algorithms, Gaussian Mixture Model Regression (GMMR). The GMMR provided the highest performance amongst the other algorithms. Stochastics models were developed to provide feature ranking for appliances energy use prediction in a low energy house Candanedo et al. (2017)[S30], the algorithms encompassed multiple linear regression, SVM with the radial kernel, GBM, random forest. Findings of this study include the identification of a relationship between weather and appliances use which presumably relates to the increase of occupancy with certain weather conditions (rain, snow, wind), and consequently increased the use of appliances. Moreover, zones such as the kitchen, laundry room, living room and bathrooms have the highest contribution in the use of appliances. This highly depends on occupancy, the function of space, and the type of appliance in the identified zones. In terms of the prediction accuracy, The GBM and RF models proved to have better accuracy when compared to the SVM-radial and multiple linear regression. One limitation identified in this study is the size of the data set used which is derived from a single case study.

Mahdavi et al. (2016)[S31] provided a simplified and stochastic model based on linear regression and other machine learning algorithms to study the relationship between occupancy presence patterns and plug loads in office buildings. The stochastic model was compared to the simplified model. The results showed that the stochastic model had better performance in terms of plug loads peaks and distribution. The limitation of this study also lied on the data availability and limited set of empirical data obtained from one case study. Also, neural network for short term appliances load forecasting in different houses is employed for the development of knowledge and data-driven model to forecast the chance of the use of a particular appliance during a given period. Historical data concerning the energy usage of different appliances and their past consumption is considered as a part of input data (Basu et al.,

2013)[S32]. The results showed that lighting appliances had the highest predictability amongst other appliances. Moreover, a rule mining clustering algorithm is utilized to distinguish major associations between energy consumption and use of appliances; users annotated activities (cooking, working..), time of day, and day of the week (Rollins & Banerjee, 2014)[S33]. This study was able to derive important associations between the aforementioned factors which promote energy saving and optimisation.

Table 2.4 Studies of Appliances Use

Study	Reference	Purpose	procedure	ML Method	Building type	Input data	Output data	Results
S29	(Wang et al., 2018)	developing models for predicting building energy use	data-driven models	multiple linear regression, adaptive linear filter algorithms (least mean square (LMS), normalized least mean square (NLMS), and recursive least square (RLS)), and Gaussian mixture model regression (GMMR)		synthetic large-size office building from DOE reference building models, the other building is an existing office building located in Des Moines, Iowa.	predict hourly energy usages in two buildings	The GMMR models outperform the adaptive filter methods
S30	(Candanedo et al., 2017)	feature ranking for the appliances energy use prediction	Stochastic models	(a) multiple linear regression, (b) support vector machine with radial kernel, (c) random forest and (d) gradient boosting machines (GBM).	low-energy house	measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station and recorded energy use of lighting fixtures.	different relationships between parameters	The GBM and RF models improved the prediction compared to the SVM-radial and multiple linear regression.
S31	(Mahdavi et al., 2016)	Studying Relationship between occupants' presence patterns and plug loads	Simplified and stochastic models	Linear regression, others	office buildings	long-term observational data monitored ccupancy, presence, plug loads.	relationship between inhabitants' presence, installed power for equipment, and the resulting electrical energy use.	The stochastic model performed better in terms of plug loads' peak and distribution.
S32	(Basu et al., 2013)	to forecast if a particular appliance will start during a given hour or not.	Knowledge and data driven model	Neural networks for short-term load forecasting	Different houses	historical data concerning the energy usage of different appliances, past consumption.	appliance usage prediction	he prediction with decision tables for lighting usage in dwellings gives the highest accuracy in all the tested cases.

S33	(Rollins & Banerjee, 2014)	identify significant associations between energy usage and four key features: hour of the day, day of the week, use of other appliances in the home, and user-supplied annotations of activities such as working or cooking.	Cluster analysis	rule mining algorithm (DBSCAN clustering algorithm)	six homes across the United States.	raw energy consumption data from several devices in each home and uses a novel *in situ* approach for soliciting user annotations to describe activities performed.	Associations Impacting Energy Consumption	while time-based associations are observed most frequently, associations between devices are common and often stronger than time-based associations

2.5.4 Shades, Blinds and Lighting Control

The modelling and prediction of occupants control of shades and blinds as well as the use of lighting were one of the first occupant behavioural controls explored, with an attempt to provide energy saving options by optimising lighting systems and their use. Figure 2.11 shows the lack of diversity in building types used as case studies in which office buildings dominate the overall studied cases with a staggering percentage of 84% where only one case study was a residential complex, and another was mixed-use commercial and office building.

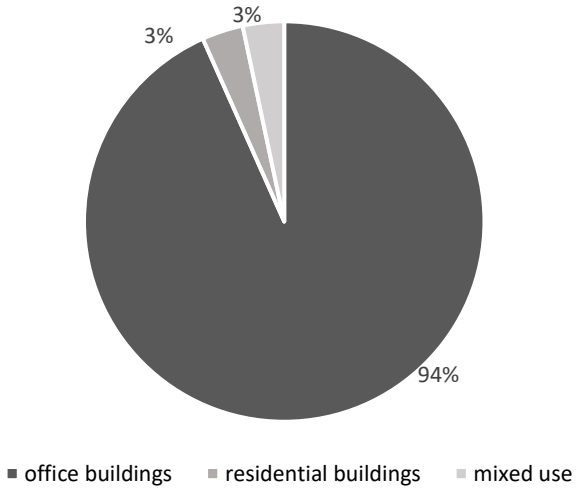


Figure 2-11 Building types

The machine learning algorithms used for the prediction of shades, blinds, or/and lighting control are of varying percentages as present in Figure 2.12. Once again, regression models were the most common models used for the predictions with a percentage of 67% while the rest of the algorithms combined formed a percentage of 33%. A review of shades, blinds and lighting control approaches by employing advanced statistical and machine learning techniques is presented in Table 2.5.

A Bayesian modelling approach developed by Sadeghi et al. (2017)[S34] attempted to model the interaction of private offices occupants with shading and electrical lighting systems. The model employed Bayesian discrete regression using individual occupant's characteristics and attributes governing occupants/shading and occupants/electric lighting interactions as part of input data.

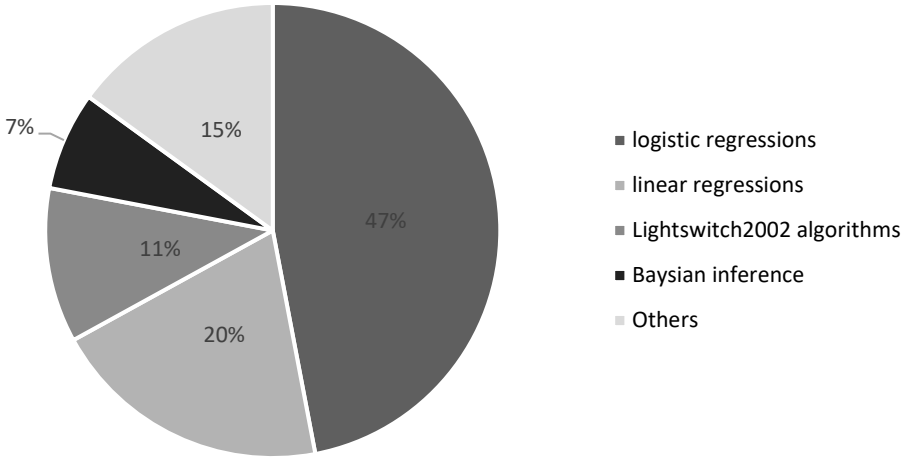


Figure 2-12 algorithms for shades, blinds and lighting control

The study suggested that occupants related attributes are significant indicators of occupants shading and lighting interactions, and to expand the predictive performance of the model, the incorporation of occupant's traits as features in shading action models is recommended. Moreover, the modelling approach increased the prediction accuracy of occupant's interactions with shading and electrical lighting in building performance simulation tool. The shortcomings of the model lied upon the limited amount of data, the uncertainty resulting from it, and the validation process. To improve the understanding of occupant's behaviour and their manual control of lighting and shading, multiple linear regressions and logistic regressions analysis were set in place in 8 single occupied offices (da Silva et al., 2013)[S35]. The analysis showed that electric lighting and shading control were influenced more by occupational dynamics i.e., occupants' arrival and departure, than by the environmental conditions. Shading and lighting deployment patterns and events probability were predicted. Inconveniency of associating results with occupants shading and lighting control patterns were present when using other case studies or experimental conditions. In another study, personalized visual satisfaction profiles in private daylit offices were inferred using a Bayesian approach (Xiong et al., 2018)[S36]. The results showed reliable visual satisfaction profiles with predicted uncertainty based on the Bayesian model's performance.

Table 2.5 Studies of shades, blinds and lighting control

Study	Reference	Purpose	Procedure	ML Method	Building type	Input data	Output data	Results
S34	(Sadeghi et al., 2017)	Modeling human interactions with shading and electric lighting systems	Bayesian Model	Bayesian discrete regression	private offices with motorized roller shades and dimmable electric lights.	dataset from a field study; environmental parameters as well as individual characteristics and human attributes governing human-shading and – electric lighting interactions.	Prediction of interactions with Shading and lighting	besides environmental variables, human attributes are significant predictors of human interactions, and improve the predictive performance when incorporated as features in shading action models.
S35	(da Silva et al., 2013)	to further our understanding of occupants' behaviour regarding the manual control of electric lighting in combination with shading control.	regression analysis	A linear multiple regression, logistic regressions	Eight single-occupied offices	environmental variables including workplane illuminance, window and background luminance and transmitted solar radiation, occupancy, occupant characteristics	predict shading deployment patterns, shading events probability	electric lighting and shading control were influenced more by occupational dynamics (arrival and departure) than by the environmental conditions
S36	(Xiong et al., 2018)	developing personalized visual satisfaction profiles	Bayesian approach	Bayesian inference	private daylit offices	Personalized visual satisfaction profiles derived from comparative preferences	Prediction of different personalized visual satisfaction profiles	Model performance results show reliable profiles with predicted uncertainty.
S37	(Fabi et al., 2014)	to describe occupants' switching on-off control over lighting	probabilistic models	multivariate logistic regression	three different office rooms	based on measurements of indoor climate parameters, outdoor environmental conditions and artificial lights "switch on/off" actions	light-switch behaviour	Two predictive light-switch behaviour models were inferred in relation to the number of actions carried out by the users (active or passive)
S38	(Yao, 2014)	to determine the impact of the control behaviour of solar shades on energy performance.	stochastic model for manual solar shades, Co-simulation with Energy plus	Logistic regression	a typical office building with internal roller shades in hot summer and cold winter zone	field measurements and logit analysis	factor s in driving solar shading adjustment	Previous studies on manual solar shades may overestimate energy savings.

S39	(Gunay et al., 2017)	Analyze the light-switch and blinds use behaviours	Behavioural model	discrete-time Markov logistic regression	ten private offices	concurrent solar irradiance, ceiling illuminance, and occupancy data	blind closing and light switching actions	the use of an adaptive lighting and blinds control algorithm can substantially reduce the lighting loads in office buildings – without adversely affecting the occupant comfort
S40	(Haldi et al., 2017)	predict the scope and effects of behavioural diversity regarding building occupant actions on window openings, shading devices and lighting.	generalised linear mixed models	Linear regressions	an office building in Switzerland and residential units in Germany and Denmark	behavioural data from three long-term monitoring campaigns	Prediction of building occupant actions	
S41	(Huchuk et al., 2016)		simplified model-based predictive control		a south-facing perimeter office space in Ottawa, Canada.			
S42	(Zhou et al., 2015)	generate lighting schedules as inputs to building simulation	Stochastic models	Logistic regression	15 office buildings	measured data occupancy, lighting pattern, indoor and outdoor conditions	lighting consumption schedule	ighting energy use was mainly driven by the occupant schedule.
S43	(Zhang & Barrett, 2012)	Studying Window blind control	stochastic model	Single variable Linear and logistic regression	naturally ventilated office building in Sheffield's, UK	A field study of occupants' window opening behaviour concerned with daily windows positioning, indoor and outdoor climatic data	the probability of windows being open given the outdoor temperature	manual window control, as indicated by the proportion of windows opened, has a strong correlation with outdoor air temperature, the season of year, time of a day and occupancy pattern. Also, windows orientation
S44	(Parys et al., 2011)	Model occupancy, use of shading system, window operation, control of artificial lighting, heat gains by	behavioural model coupled with building simulation smodel	Logistic regression	Office buildings			

		appliances and the control of heating and cooling setpoints.					
S45	(Reinhart & Wienold, 2011)	Provide computer-based daylighting analysis	Annual daylight glare probability profiles combined with an occupant behaviour model		Office building	field study data that monitored long term occupancy and use of light switches and shading devices, climatic data	annual shading profiles and visual comfort conditions, autonomy plots, energy loads, operational energy costs and greenhouse gas emissions.
S46	(Haldi & Robinson, 2010)	prediction of the usage of shading devices	stochastic model for simulating blind usage	Logistic regression	office buildings	initial blind status, indoor and outdoor illuminance , the occupancy, thermal and visual parameters influencing actions on shading devices	Shading devices usage
S47	(Daum & Morel, 2010)	the influence of lighting and blind control models on the heating, cooling, and lighting energy loads	Stochastic models	Logistic regressions	Office room		

Moreover, predictive probabilistic models were derived to describe occupants on/off lighting control in 3 different office rooms by employing multivariate logistic regression (Fabi et al., 2014)[S37]. Only interactions of categorical and continuous variables were studied to avoid modelling complexities. The study deduced that there is a negative correlation between the probability of switching on the ceiling lamps and the room temperature, sun elevation, and daylight coming from windows. Another relevant study presented a co-simulation of the stochastic model for manual solar shades with EnergyPlus (Yao, 2014)[S38]. The stochastic model relied on logistic regressions and proved more reliability while confirming the limitations and overestimation of previous studies on manual solar shades. A noteworthy study is based on a behavioural model derived from adaptive lighting and blind control algorithms designed to examine the blinds and light switch occupants control behaviour in ten private offices (Gunay et al., 2017)[S39]. The results demonstrate that the adaptive light and blinds control algorithms can be utilised to provide a substantial reduction in energy use derived from lighting loads while adjusting and responding to the occupant's visual comfort. The studies S40-S47 are explained in Table 2.5.

2.6 Energy Prediction Accuracy of Machine Learning Techniques

This section provides the details of energy prediction accuracy of machine learning techniques shown in Table 2.6. Fayaz and Kim (2018) compared ANFIS (adaptive neuro-fuzzy inference system) and ANN (artificial neural network) and the dataset has been gathered from four multi storied residential building. The authors addressed the proper utilization of energy which is a lot wasted annually. The proposed model is evaluated using MAE, RMSE and MAPE evaluation metrics. The results revealed that ANFIS outperform ANN model in prediction accuracy. Truong et al. (2021) compared XGB (extreme gradient boost), MLR (multiple linear regression), and Shallow ANN for a better result. The dataset was created using synthetically data (six household residential building data) and the Deep ANN Machine learning approach. The accomplishment the system's technological impediments to accurate load forecasting were addressed with the goal of enabling larger adoption of the decentralized generation model and better decision-making.

Due to reduced efficiency caused by substantial transmission losses and waste energy, the centralized electricity producing model has several disadvantages for the environment and end-user. It may be used to optimize learning rate, momentum, iteration, and batch size, which are all hyper-parameters related to ANN models. Parhizkar et al. (2021) compared LR (linear regression), SVR (support vector machine), RT (regression tree), RF (random forest), and KNN (k nearest neighbor) for a better accuracy. The dataset was created using meteorological data (gathered over the last four years from residential buildings and four different datasets) and deep machine learning. The technique of preprocessing (principal component analysis) was utilized. They tackled the achievement of removing noisy features in conjunction with a prediction approach for improved and more effective decision-making. Climate change, building structure, occupation, and geographic location all contribute to energy waste.

Lei et al. (2021) compared BPNN (back propagation neural network), ENN (elman neural network), FNN (fuzzy neural network), and RS-DBN (rough set mixed with deep belief neural network) for a superior result. The dataset included in the paper was compiled from a variety of sources. Rough set reduction data was collected from 100 civil public buildings, as well as data from a university's laboratory building, and a deep learning method was employed in conjunction with a rough set theory technique. The accomplishment improves energy system control and utilization through precise prediction, as well as lessen influence factors, which they addressed. Inadequacy of physical-model-based and statistical approaches, ML method, and energy waste previous research was lacking, and the data was limited.

Table 2.6 Energy prediction accuracy of machine learning techniques

Authors	Technique	Algorithm	Prediction Accuracy						
			MAE	RMSE	MAPE	MRE	R^2	MAD	CV
Fayaz and Kim (2018)	Deep Extreme Learning Machine	ANFIS	2.45	2.81	6.12	X	X	X	X
		ANN	2.43	4.85	7.08	X	X	X	X
Truong et al. (2021)	Deep ANN Machine learning	Deep ANN	X	111.20	X	X	97.5	X	X
		XGB	X	270.85	X	X	84.9	X	X
		MLR	X	634.65	X	X	17.2	X	X
		Shallow ANN	X	636.76	X	X	16.6	X	X
(Parhizkar et al., 2021)	Deep machine learning (Principal component analysis)	RF(Te)	X	0.23	X	X	0.9	X	X
		RT(Ta)	X	1.02	X	X	0.9	X	X
		RF(Ya)	X	0.21	X	X	0.9	X	X
		RT(BA)	X	1	X	X	0.9	X	X
(Lei et al., 2021)	Deep learning algorithm and integrated with rough set theory	DBN	X	0.05	0.05	X	X	X	X
		RS-DBN	X	0.03	0.03	X	X	X	X
		DBN	X	0.04	0.04	X	X	X	X
		RS-DBN	X	0.02	0.02	X	X	X	X
(Nie et al., 2021)	GBRT	RNN	-29.4	1.4	X	X	X	X	X
		SVM	89.4	70.5	X	X	X	X	X
		ARIMA	87.4	92.9	X	X	X	X	X
		ARIMA-GBRT	96.1	91.4	X	X	X	X	X
		ARIMA-RNN	96.5	92.1	X	X	X	X	X
(Kim & Cho, 2019b)	Deep Learning CNN-LSTM	LR	0.502	0.651	83.7	X	X	X	X
		LSTM	0.526	0.717	44.3	X	X	X	X
		CNN-LSTM	0.331	0.595	32.8	X	X	X	X
		LR	0.319	0.384	41.3	X	X	X	X
		LSTM	0.243	0.323	35.7	X	X	X	X
		CNN-LSTM	0.238	0.308	31.8	X	X	X	X
Le et al. (2019)	Deep Learning CNN & Bi-LSTM	LR	0.502	0.652	83.7	X	X	X	X
		LSTM	0.526	0.717	44.3	X	X	X	X
		CNN-LSTM	0.332	0.596	32.8	X	X	X	X
		EECP-CBL	0.392	0.546	50.0	X	X	X	X
		LR	0.320	0.385	41.3	X	X	X	X
		LSTM	0.244	0.324	35.7	X	X	X	X
		CNN-LSTM	0.238	0.309	31.8	X	X	X	X
		EECP-CBL	0.177	0.220	21.2	X	X	X	X
Wen et al. (2020)	Deep Learning (DRNN-GRU)	DRNN-GRU	0.34	0.51	3.50	X	X	X	X
		DRNN-LSTM	0.39	0.56	3.64	X	X	X	X
		Deep RNN	0.94	1.11	8.59	X	X	X	X
		MLP	1.67	2.20	14.4	X	X	X	X
		ARIMA	1.70	2.25	15.0	X	X	X	X
		SVM	2.14	2.84	32.7	X	X	X	X
		MLR	4.60	6.07	37.7	X	X	X	X
	Deep Learning	FNN	0.70	1.05	X	X	0.47	X	X
		DFNN	0.55	0.84	X	X	0.53	X	X

(Kiprijanovska et al., 2020)		TCN	0.50	0.78	X	X	0.59	X	X
		LSTM	0.54	0.81	X	X	0.54	X	X
		GRU	0.54	0.80	X	X	0.54	X	X
		HousEEC	0.23	0.44	X	X	0.90	X	X
Dong et al. (2021)	Ensemble energy and pattern classification	SPD	X	4.44	X	X	X	X	0.06
		SVR	X	7.35	X	X	X	X	0.09
		ANN	X	5.71	X	X	X	X	0.10
Syed et al. (2021)	Novel Hybrid Deep Learning model	LR	85.5	137.2	70.3	X	X	X	0.10
		ELM	53.4	90.1	65.8	X	X	X	0.16
		LSTM	4.3	6.33	2.5	X	X	X	0.99
		DL-LSTM	3.4	5.44	2.0	X	X	X	0.8
Peng et al. (2021)	EDA-LSTM	SVR	22.0	24	17.9	X	X	X	X
		RFR	24.0	27.0	19.3	X	X	X	X
		AdaBoostR	28.0	31.0	22.2	X	X	X	X
		LSTM	38.0	48.0	5.6	X	X	X	X
		Dual A-LSTM	32.0	41.0	4.1	X	X	X	X
		EDA-LSTM	26.0	31.0s	4.3	X	X	X	X
(Somu et al., 2020)	ISCOA-LSTM	ARIMA	0.37	0.48	X	X	X	X	X
		DBNR	0.32	0.41	X	X	X	X	X
		SVR	0.14	0.15	X	X	X	X	X
		GA-LSTM	0.32	0.14	X	X	X	X	X
		PSO-LSTM	0.42	0.09	X	X	X	X	X
		SCA-LSTM	0.44	0.46	X	X	X	X	X
		ISCOA-LSTM	0.03	0.05	X	X	X	X	X
Kim and Cho (2019b)	CNN-LSTM neural network	LSTM	0.6	0.8	X	51.4	X	X	X
		GRU	0.6	0.8	X	51.3	X	X	X
		Bi-LSTM	0.5	0.8	X	51.1	X	X	X
		CNN-LSTM	0.3	0.6	X	34.8	X	X	X
Jamil et al. (2021)	Block chain deep learning peer-to-peer	RNN	422.2	567.5	2.94	X	0.9	X	X
		LSTM	377.2	519.9	2.6	X	0.9	X	X
		RF	1328.2	1064.2	14.7	X	0.4	X	X
		XGBoost	943.3	793.1	9.91	X	0.5	X	X
X.-B. Jin et al. (2021)	Attention based Encoder-Decoder with Bayesian optimization (GRU-LSTM)	RNN	512.6	613.2	X	X	X	X	X
		LSTM	498.8	600.4	X	X	X	X	X
		GRU	487.9	586.8	X	X	X	X	X
		GRU-LSTM	458.9	550.3	X	X	X	X	X

Nie et al. (2021) developed a unique energy consumption prediction model to simulate and predict energy use with greater accuracy. The wasting of a large amount of energy has a negative impact on the environment. The dataset was collected from a residential residence, and GBRT (gradient boosting regression tree) was utilized to modify GB using the RT of fixed size

technique. Kim and Cho (2019b) compared LSTM (long short-term memory), LR (linear regression), and CNN-LSTM for a superior result. LSTM is utilized as a classifier, and CNN is used to extract complicated characteristics from images. The dataset was collected from collected data on energy usage from residential houses over a four-year period, and Deep learning CNN-LSTM approach was employed. They addressed the achievement of minimizing energy waste and economic loss, as well as effective forecast. There is a lot of energy waste from various places and for various causes. They need to collect energy consumption from a larger number of houses to confirm it, as well as other factors such as occupancy and major effects.

Le et al. (2019) compared LR (linear regression), LSTM (long short-term memory), EECP-CBL (electric energy consumption prediction model utilizing the combination of CNN and Bi-LSTM) to provide a better result. A CNN Bi-LSTM is a bidirectional LSTM and CNN architecture that is combined. It learns both character level and word level characteristics in the original formulation for named entity recognition. The dataset was obtained from the IHEPC (individual household electric power consumption dataset) dataset. Data on energy use from the previous five years was collected, and Deep Learning CNN and Bi-LSTM techniques were applied. This cuts down on energy waste and economic disruption while also allowing for more precise predictions. They addressed these issues to make better decisions about how to save energy. For a variety of causes, a lot of energy is squandered. The waste of a significant amount of energy has a negative economic impact. The results showed that several strategies, such as evolutionary algorithms and optimized models, must be used to increase performance. Wen et al. (2020) compared with DRNN-GRU is done to see whose results are better Deep RNN, MLP, ARIM, SVM, and DRNN-LSTM. The dataset for the paper was compiled from The Deep Learning (DRNN-GRU) technique was utilized to gather the load demand data for residential structures from the Data port website. Achieve excellent forecasting accuracy with few input variables by accounting for time dependencies. The issue is the rise in daily load demand and energy consumption. The drawback is that deeper networks and finer-grained data might potentially be used. An LSTM-based neural network, Pecan Street, and

deep learning techniques are employed (Kiprijanovska et al., 2020). Extensive datasets for household electricity consumption were gathered. Accurate load forecasting results are provided by the proposed model. False closest neighbor algorithm (FNN), deep feedforward neural network (DFNN), temporal convolutional network (TCN), long short-term memory (LSTM), gated recurrent unit (GRU), and houseEEC are compared. To creates a new classifier that performs better than any of its constituent classifiers, ensemble learning generates a variety of basic classifiers. These base classifiers may vary in terms of the training data, representation, or the algorithm being employed. Dong et al. (2021) used ensemble energy and pattern categorization. The achievement is each prediction model has the capacity to manage a single pattern and provide more precise predictions. The authors compared the best results of SVR and ANN for SPD (Stacking Pattern Decision).

Syed et al. (2021) proposed a New Deep Learning Framework for Classifying Hyperspectral images using subspace-Based Feature Extraction and Convolutional neural networks and compared LR, ELM, & LSTM. Eight textual tweets and review datasets of various disciplines are used to build and test hybrid deep sentiment analysis learning models that integrate long-term memory networks, convolutional neural networks, and support vector machines. The dataset utilized was compiled from information gathered from residential buildings using a novel hybrid deep learning model approach. Enhancing load and scheduling prediction accuracy is the accomplishment. The issue is overusing energy-consuming equipment wastes a significant portion of the energy used in buildings. The restriction is that training an energy forecasting model might potentially be done in parallel with bidirectional LSTMs to accomplish distributed computing. Peng et al. (2021) compared SVR RFR, AdaBoost Regression, and LSTM with the purpose of ensuring the best results. The dataset analyze three years' worth of daily electricity consumption data from residential buildings in Shanghai's Pudong neighborhood. (Somu et al., 2020) compared ARMIA (auto regressive integrated moving average) SV regression, BN regression to get the best outcome.

Jamil et al. (2021) compared RNN, LSTM, RF, and XGBoost to see which produces the best results. The dataset used was compiled using the

block-chain deep learning (peer-to-peer) technique with data from the Korean province of Jeju. The accomplishment is the intelligent peer-to-peer energy trading, data analysis, and predictive analysis supported by smart contracts.

Energy consumption is increasing in today's world due to the sharp increase in human population and technological development. Hence more accurate prediction of energy consumption is important. Kim and Cho (2019b) proposed a hybrid model based on CNN and LSTM. The authors used CNN for theextraction of complex features from multiple variables and LSTM for modelling irregular time series data. Using the proposed model this study achieves better results as compared to previous studies in this area and records a small value of root mean square error. Pham et al. (2020) presented a proposed random forest (RF) based model for short-term energy consumption prediction. Five one-year datasets are used in this study for analysis of models. Models such as the RF model, M5P, and Random Tree (RT) are compared for comparativeanalysis. For evaluation of the results, accuracy measures are used such as MAE, MAPE, RMSE,and SI. The experimental results indicated that the RF model has a better prediction accuracy in the prediction.

Divina et al. (2020) proposed a neuro-evaluation-based approach using genetic algorithms to find the optimal set of hyper parameters to configure the deep neural network. This method is then used to forecast the electrical energy consumption. The right set of hyper parameters is significantly helpful to deep neural networks for excellent performance. The proposed model performance is then assessed by experimenting using a large dataset of 10 years in Spain. The results are obtained using MRE and SD as accuracy measures. After comparing the results with NDL, CNN, LSTM, FFNN, ARIMA, DT, GBM, RF, EV, NN, and ENSEMBLE, the authors claimed that the methodology they proposed has the capability for short-term electric energy prediction, and on the specific dataset that is used achieved the top performances.

Chitalia et al. (2020) proposed a framework for short-term energy prediction to capture non-linearity. The author explored nine multiple hybrid neural networks and clusters. Data of five commercial buildings are combined

from five multiple places in Bangkok Thailand, Hyderabad-India, Virginia-USA, New York-USA, and Massachusetts-USA. The author RMSE, MAPE, and CV as an accuracy measure. The resultsproved that the deep learning algorithms provide 20–45%perfection in energy prediction performance in a comparison with other conventional models for both hour-ahead and 24-ahead load prediction.

Fekri et al. (2021) proposed an online Adaptive RNN has the capability of learning freshly arriving data continuously and updating weights of RNN according to the current data. The new method is evaluated on the real-world data taken from five household customers provided by London Hydro. LP, linear, passive-aggressive, bagging, and KNN regression, RNN are used for comparative analysis using MSE and MAE as accuracy measures. The results proved that the proposed method achieved higher accuracy than the traditional offline long short-term memory network and five other online algorithms.

Kim and Cho (2019a) predicted residential energy consumption using artificial intelligence neural networks which optimize using trial and error operators which lack prior knowledge. For comparative analysis of model's household, the UCI repository is used for datasets collection. The proposed model achieves the best prediction of the results in terms of accuracy and lowermost mean squareerror (MSE) as compared to conventional models.

(He et al., 2019) extracted important features from external factors affecting the energy use forecasting by LASSO regression. The author then developed a LASSO-QRNN model to predict annual electricity consumption. The model was evaluated using MAP E and RMSE showing precision in prediction.

The results of machine learning models' energy accuracy consumption prediction are elaborated in following Table 2.7.

Table 2.7 Prediction accuracy results of machine learning techniques

Studies	Technique	MSE		RMSE		MAE		MAPE	
		LT	ST	LT	ST	LT	ST	LT	ST
(Le et al., 2019)	EECP-CBL	0.29	0.05	0.54	0.22	0.39	0.98	50.09	11.66
Divina et al. (2020)	NDL	1.4	3.01	×	×	×	×	×	×
Pham et al. (2020)	RF	×	×	5.18	2.2	3.4	1.47	23.52	9.3
Chitalia et al. (2020)	LSTM	×	×	6.35	2.47	×	×	74.51	19.22
Fekri et al. (2021)	RNN+LSTM	66.98	28.82	×	×	21.18	11.88	×	×
Kim and Cho (2019a)	PSO-based CNN-LSTM	0.39	0.44	×	×	×	×	×	×
(He et al., 2019)	Lasso-QRNN	×	×	×	2.04	×	×	×	0.05

2.7 Summary

This chapter aimed to establish a better understanding of how occupant's behaviour can be better predicted through machine learning approaches to improve building energy predictions. In this chapter, machine learning approaches have been reviewed with the goal of their application in advancing occupant behaviour modelling and prediction. Due to the occupant's behaviour stochastic nature, in most cases, it is difficult to extrapolate building occupant related parameters. Energy simulation approaches alone without the consideration of the impact of occupant's behaviour may fail to obtain accurate simulations and predictions. Studying and predicting occupant energy behaviour is complex and presents non-linearity in data, patterns of behaviour, dependencies, relationships and constraints, hence in most cases data is over or under fitted. However, machine learning methods can deal with non-linearity and provide more reliable outcomes. It is evidenced that the machine learning approach can extrapolate valid, insightful, and inclusive building occupant behaviour patterns.

This chapter concludes the following advantages and limitations in terms of occupant's behaviour prediction employing machine learning approaches.

i. Machine learning approaches have proven to provide more reliable predictions when compared with the deterministic approaches with accuracy levels reflected throughout the review sections. It provides a better understanding of occupant's behaviour stochastic nature and incorporation of its impact on building energy performance. This all contributes to better building energy performance prediction and minimising the performance gap.

ii. The key limitations are represented by shortcomings in the model validation process, and the generality of estimate parameters in which some variables are based on estimated values instead of the actual values due to the lack of adequate datasets. This, in turn, affects the validity of the models.

iii. As evidenced in the review, some cases presented the difficulty of relating machine learning modelling results with occupant behavioural control patterns derived from other case studies or different conditional input process. For example, a predictive model applied for residential buildings requires different inputs (conditional and non-conditional) to school buildings. Also, this relates to the lack of state-of-the-art predictions on different types of case studies which is mainly limited to residential and office building as evidenced by the reviewed papers where these buildings represented over 90% of the case studies.

iv. A high level of expertise is needed to achieve an accuracy as machine learning approaches can be complicated by nature, and a high level of knowledge is needed to develop the model.

AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE:
CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

Research Methodology
Chapter 3

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this research, the research questions are answered using multiple research methods. It can, therefore, be regarded as a mix research method. Research methods are selected based on their suitability to answer each research question. The first research question RQ1 is related to the investigation of energy consumption prediction models and accuracy improvement in ensemble and solo machine learning techniques; the second research question RQ2 is about the development of the occupant's behaviour-based machine learning model to improve energy consumption accuracy prediction of residential buildings; the RQ3 and RQ4 are associated with the evaluation and applicability of the proposed model respectively. Table 3.1 depicts the research methodology used in this research.

Table 3-1 Mapping of research methods with chapters

Research Methods	2	3	4	5	6	7
	Literature Review	Methodology	Model Development	Data Collection & Evaluation	Results & Discussion	Conclusion & Recommendations
Literature Review (Secondary study)	✓					
Engineering			✓			
Experiments				✓	✓	
Comparison Evaluation performance metrics					✓	

3.2 **Research Design and Approach**

There are four categories of conducting research such as scientific, engineering, empirical and analytical (Hasselbring & Giesecke, 2006). In the first stage of the scientific method, the problem needs to be observed thoroughly; then a model shall be proposed. In the second stage, the proposed model shall be validated using formal methods and theory to prove the hypotheses. In the last stage, this process will eventually be repeated as far as possible.

Instead, the engineering methodology involves analyzing existing solutions to gain knowledge on how to develop a better solution and testing it to verify hypotheses. In the empirical method, the first step is to formulate a model. The statistical or qualitative methods must be established accordingly to validate the given hypotheses. In the final stage, the planned model must be applied to case-study for evaluation purpose. Alternatively, the methodology of an analytical method is different; in this case, first, a formal theory is formulated then the proposed theory is established to derive results. These results are then compared with empirical observation if possible.

Looking at the above categories, this study has selected the engineering method. The primary focus of supporting the engineering methodology is to prioritize the design and construction of a more efficient solution than the current solutions available. Therefore, the engineering method is the most suitable research method to be applied to this research as the main purpose of this research is to improve the accuracy prediction of energy consumption of residential buildings. The engineering method contains four steps which are (1) observe the existing solutions; (2) propose a new solution; (3) develop the proposed solution; and (4) measure an analyze as shown in Figure 3.1.

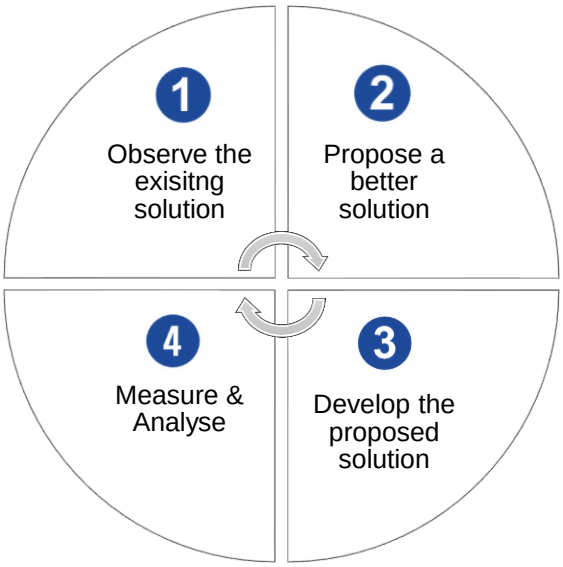


Figure 3-1 Engineering research method process

In the first step, the prior studies on energy consumption prediction models and accuracy improvement in ensemble and solo machine learning techniques in residential buildings are reviewed. As part of this study, existing models, methods, techniques, and frameworks for predicting energy consumption accuracy were analyzed. After this analysis, an ensemble machine learning model based on occupancy behavior was proposed in order to improve the accuracy of energy consumption prediction for residential buildings. In the third step, the idea behind the proposed solution is to incorporate machine learning techniques to make an ensemble model. The model development stage is discussed in chapter 4. In the last step, the solution is evaluated using evaluation metrics and real case studyand if it fails to achieve the objectives of this research, these four steps shall be repeated to find a better solution for the problems in this research.

This study utilized the research plan guidelines proposed by Wohlin and Aurum (2015) to design its research structure. The research plan was structured into three phases: strategy, tactical, and operational, with a total of eight decision points, as illustrated in Figure 3.2. Various methods were available for executing each decision point.

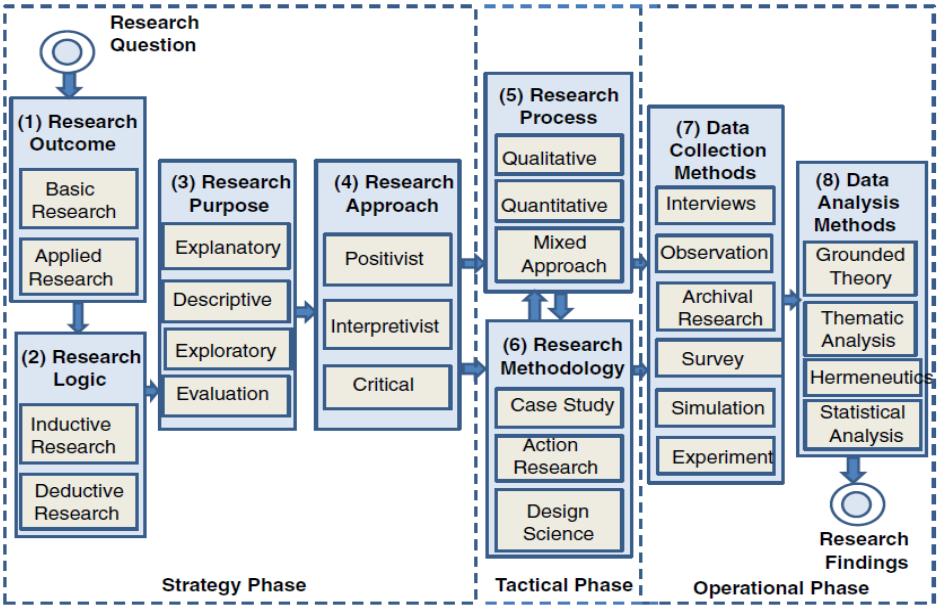


Figure 3-2 Structure of research phases. Wohlin and Aurum (2015)

This research applied the guidelines to map the decision making procedure of the research design onto the research structure illustrated in Figure 3.3.

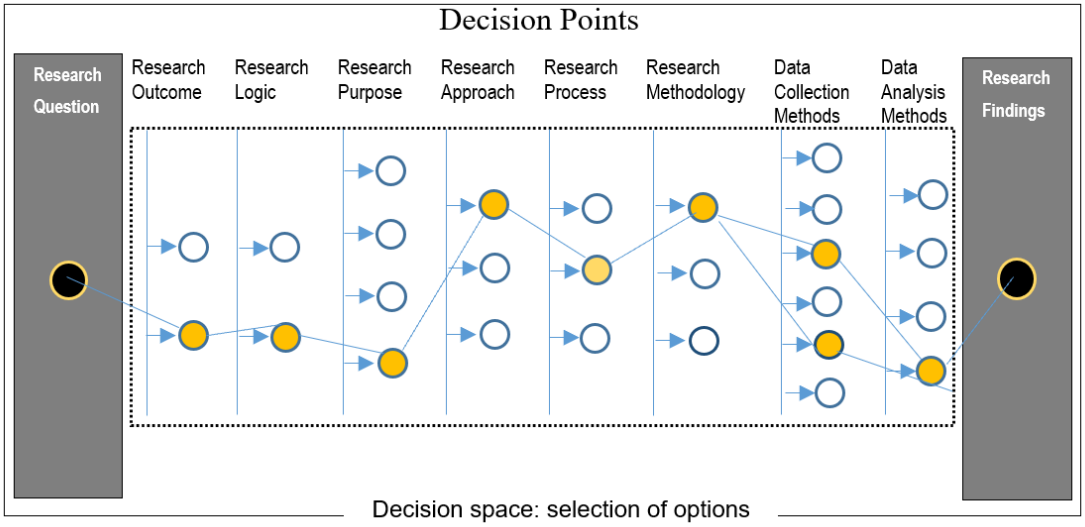


Figure 3-3 The procedure of research decision making

3.2.1 Research Strategy Phase

Outcome: This research aims to propose a novel ensemble machine learning model for improving the accuracy prediction of energy consumption of residential buildings. The proposed model aims to solve a specific problem of

residential buildings accuracy prediction. Thus, the research outcome of this study is based on applied research.

Logic: In this research, the deductive method was employed for research logic. The deductive reasoning involves moving from general to specific in a top-down approach. It begins with a theory, develops hypotheses from that theory, and then collects and analyzes data to test those hypotheses. This research followed the same approach by first identifying the problem through a thorough literature review, establishing a hypothesis regarding energy consumption accuracy prediction, and evaluating it after collecting and analyzing data. The hypothesis that 'the application of machine learning techniques can provide more accurate energy predictions and reliable occupant behaviour considerations' was tested by employing quantitative methods to confirm it.

Purpose: The primary objective of this research is to create a machine learning method for energy performance prediction of buildings that considers and incorporates all factors affecting occupant behavior. The main focus is to improve the prediction accuracy compared to existing models by including the impact of occupant behavior on energy consumption. The goal is to reduce the energy performance gap and provide more reliable predictions.

Approach: This research has adopted the positivist approach, which is characterized by an objective and empirical approach to obtain accurate and reliable results that can be replicated. The role of the researcher is limited to the collection and analysis of "objective" data to achieve quantifiable outcomes.

3.2.2 Research Tactical Phase

Research Process: For the research process decision point, a quantitative method was chosen as the research focuses on the accuracy prediction of energy consumption in residential buildings, which requires various metrics for

evaluation and statistical techniques to analyze data. As for the research methodology, a case study approach was utilized as it allows for multiple analyses and different perspectives to be applied to the selected software projects.

Research Methodology: As for the research methodology, a case study approach was utilized as it allows for multiple analyses and different perspectives to be applied to the selected software projects.

3.2.3 Research Operational Phase

Data Collection: Datasets and case studies are used for data collection methodology. Whereas this research aims to improve the accuracy prediction of energy consumption.

Data Analysis: This research has used statistical analysis to analyze data using python statistical libraries such as NumPy, SciPy, Sci-kit learn, in order to provide descriptive analysis, factors important analysis, occupants impact analysis, regression and machine learning analysis.

3.3 Operational Framework

In this section, the plan for implementing the research objectives in each stage of the study is described. Figure 3.4 illustrates the operational framework, including all processes and sub-processes, as well as the activities conducted and the delivered outcome.

3.3.1 PHASE-1: Literature Review and Planning

In the planning phase, a thorough examination of the activities that will take place in each stage of the research was conducted. A Gantt chart that outlines the milestones and deadlines for each phase of the research was developed to ensure well-timed execution of the thesis. The Gantt chart is presented in Table 3.2. Every step of the research process was completed in

detail at this phase. The Gantt chart for research milestones was primarily created to ensure prompt research execution. A research milestone is a structured, established description of the problem that the study is attempting to solve. It facilitates research planning by dividing the activities into important subtasks. It is useful for defining the milestones and paths between them.

At this phase, every step of the research process had been meticulously finished. The main goal the Gantt chart for the research milestones was to guarantee swift research execution. An organized, formal explanation of the issues that the study is aiming to answer is a research millstone. It makes research planning easier by breaking down the tasks into significant subtasks.

The chart was utilized to specify the milestones and the routes to be taken to achieve them. The research was split into three primary stages: investigation, model development, and evaluation, with each stage having sub-tasks and planning milestones. During the literature review phase, the current state of energy consumption prediction models and the accuracy improvement in both solo and ensemble machine learning techniques were explored, which aimed to achieve the first research objective of investigating the energy consumption prediction models and accuracy improvement in ensemble and solo machine learning techniques.

This phase provided a literature review and the output of the review delivered insights to the occupant related parameters affecting energy performance. It analysed the machine learning techniques that best captures occupant's behaviour for more accurate energy predictions. This has provided the foundation to the implementation phases of the research. Understanding occupant's behaviour can yield better building energy predictions. In this phase, several machine learning approaches have been studied with the goal of advancing occupant behaviour modelling and prediction. Due to the occupant's behaviour stochastic nature, in most cases it is difficult to extrapolate building occupant related parameters from explored case studies by using conventional approaches. Conventional approaches may fail to obtain accurate simulations and prediction resulting from its limitation to handle complex and non-linear problems, and reflect the patterns of data, the dependencies, relationships and constraints, in most cases data is over or

under fitted. However, machine learning methods evidenced the ability to extrapolate valid, insightful, and inclusive building occupant behaviours patterns.

In conclusion, in terms of occupant's behaviour prediction by means machine learning algorithms the following advantages and limitations are presented:

The advantages lie in the following:

i) providing more reliable predications when compared with the conventional simulation approaches in most cases.
ii) providing a better understanding of occupant's stochastic nature and its impact.
iii) contributing to better building energy performance prediction and minimizing the performance gap.

The key limitations are summarized as follows:

i) shortcomings in the model validation process.
ii) Difficulty of relating machine learning modelling results with occupant behavioural control patterns derived from other case studies or different conditional input.
iii) The complexity and level of knowledge needed to develop the model.
iv) Cost and computational power required.
v) The lack of state of art predictions on different types of case studies.
vi) The complexity of use and interoperability of the model with other simulation models.

This phase presented an overview of machine learning approaches with an endeavour to overcome the shortcomings of the conventional prediction models by providing reliable models of energy-related behaviour with higher accuracy and reusability potential. Tables 3.2 and figure 3.4 summarises research roadmap and operational framework.

Table 3-2 Gantt chart of research milestone

No	Task	Dur	Start	Finish	2018		2019		2020		2021		2022	
					Jan-June	Jul-Dec	Jan-June	Jul-Dec	Jan-June	Jul-Dec	Jan-June	Jul-Dec	Jan-June	Jul-Dec
1	**STAGE-1 To investigate and analyze the existing energy consumption prediction models, approaches and methods**	2 Y	01-Jan-18	31-Dec-19										
	Preliminary research initiation and plan	6M	01-Jan-18	30-Jun-18										
	Literature review on existing energy consumption prediction approaches	1Y6M	01-Jul-18	31-Dec-19										
2	**STAGE-2 To design and develop ensemble energy consumption prediction model**	1Y	01-Jan-20	31-Dec-20										
	Research design, methodology and strategy	3M	01-Jan-20	31-Mar-20										
	Operational framework	2M	01-Apr-20	31-May-20										
	Development of proposed ensemble energy consumption prediction model	4M	01-Jun-20	30-Sep-20										
	Conceptual and mathematical relationship justification	3M	01-Oct-20	31-Dec-20										
3	**STAGE-3 To evaluate the proposed ensemble energy consumption prediction model**	2Y	01-Jan-21	31-Dec-22										
	Preparation of data collection	4M	01-Jan-21	30-Apr-21										
	Collect evidence	4M	01-May-21	31-Aug-21										
	Editing and coding data	4M	01-Sep-21	31-Dec-21										
	Computation of energy consumption	4M	01-Jan-22	30-Apr-22										
	Evaluation process	4M	01-May-22	31-Aug-22										
	Interpretation of results	4M	01-Sep-22	31-Dec-22										

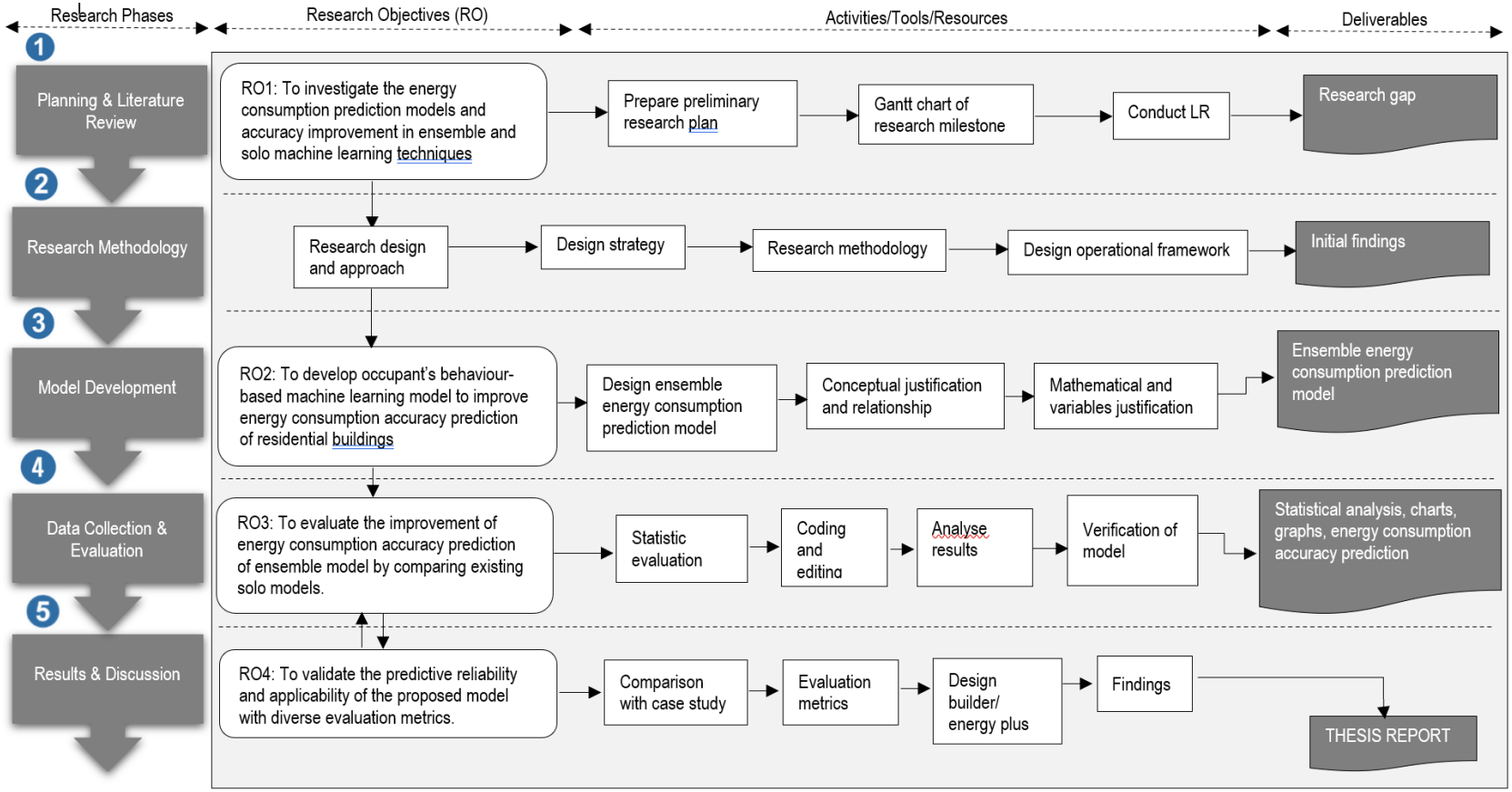


Figure 3-4 Operational Framework

3.3.2 **PHASE-2: Research Methodology**

During this phase, the research design, strategy, methodology, as well as the operational framework are developed to ensure the viability and accessibility of the research. Those phases were previously discussed in this chapter. However, the flowchart for the research methodology is presented in Figure 3.5. In addition, the design of an ensemble energy consumption prediction model is proposed at the end of this phase. A summary of the operational framework activities and outputs are presented in Table 3.3.

Table 3 -3 Phase 2 of operational framework

Phase-2		
Activities	resources	outputs
Research design	Libraries for literature review Research methodology	Engineering methodology
Research strategy	Decision making procedure	Research strategy phases
Research methodology	Flowchart	(figure 3.5) flowchart
Operational framework	Coding and simulation	Indepth development

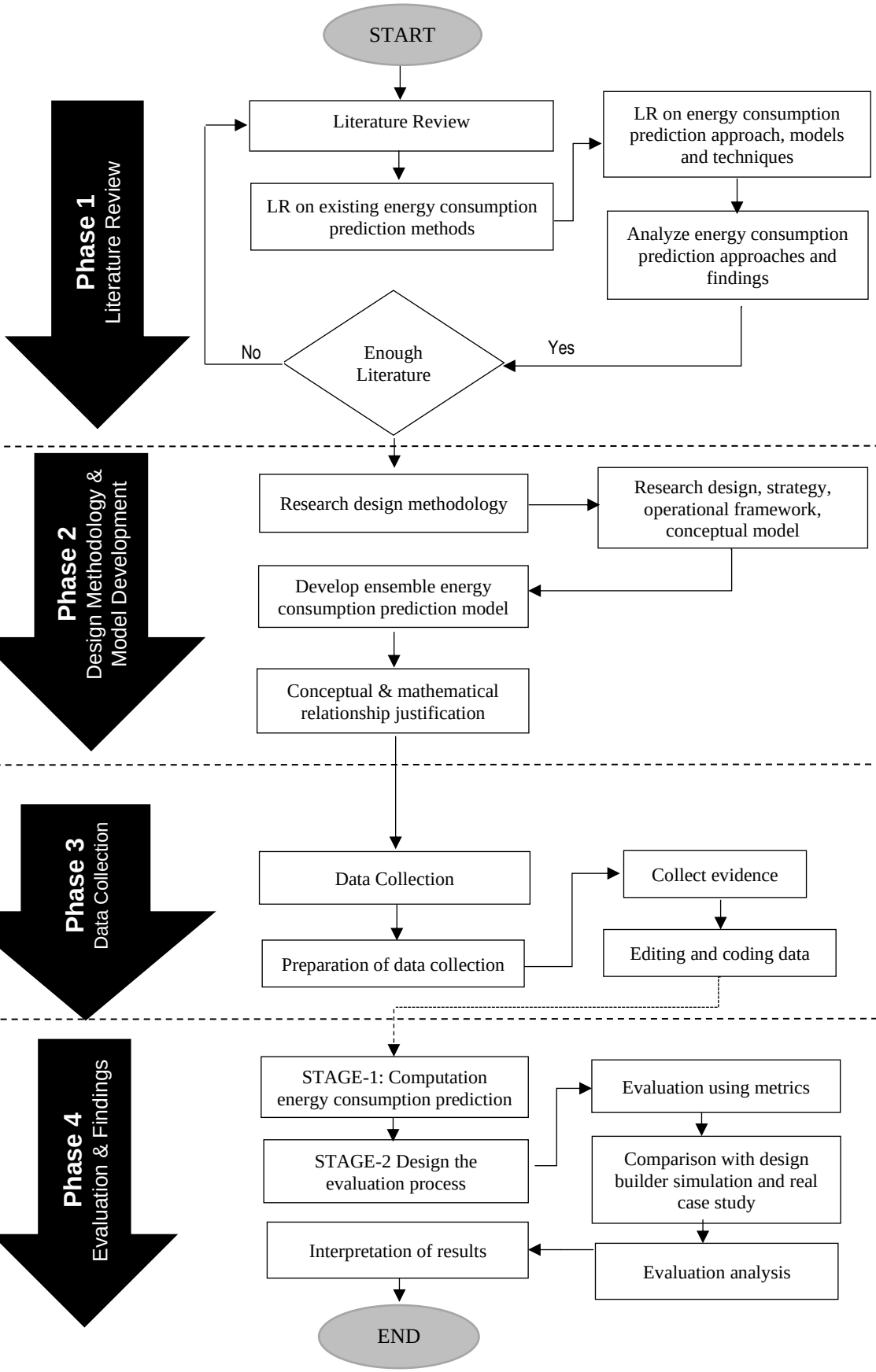


Figure 3-5 Flow chart research methodology

3.4 Methodology of Literature Review

The systematic review follows the guidelines set by Kitchenham and Charters (2007). It evaluates the machine learning techniques used to predict the impact of occupant's behaviour on building energy performance. The review investigates the machine learning techniques used to predict occupant's behaviour contribution to building energy consumption to complement the building energy model to provide accurate and more reliable result and reduce the performance gap. This research summarizes the methods and results of selected machine learning techniques in predicting occupant's behaviour and their impact on energy performance and identifies any gaps in current research to identify areas for future research.

This research review covers the following points.

i) Explores the machine learning techniques for predicting occupant's behaviour and their impact on building energy performance found in current literature,
ii) Highlights the benefits and drawbacks for identified machine learning techniques in terms of predicting occupant's behaviour on energy performance,
iii) Assesses the extent of accuracy expected for predicting the impact of occupant's behaviour when applying machine learning techniques.

3.4.1 Search Strategy

The review's search process consists of two stages. In the first stage, relevant studies were identified by searching through six commonly used digital libraries: Science Direct, IEEE, ACM Library, Web of Science, spronger , and SCOPUS. The search was limited to articles published from 2005-2022, and the search query string was tailored to each database to find studies related to the research questions depicted in Figure 3.6.

energy consumption* AND occupant's behaviour*
(energy OR consumption OR occupants OR occupancy OR behaviour) AND (estimate OR predict OR forecast OR calculate) AND (machine learning) AND (modelling OR method OR technique OR approach) AND language (English)

Figure 3-6 Search string

The literature review sheds light on the application of machine learning approach to predict occupants related parameters contribution building energy performance; it provides an appraisal of current literature to explore the adaption of the machine learning approaches in building energy performance prediction.

Figure 3.7 shows PRISMA flow diagram of the search strategy for the literature review process. A total of 317 research papers were obtained from the six digital libraries used in the first stage of the search strategy, and were downloaded for further examination. The downloaded studies were combined in Endnote version X20, and duplicates were removed, resulting in 174 studies. In the second stage, inclusion/exclusion criteria were applied by manually examining the title, abstract, introduction, and conclusion/full text of the remaining studies, which resulted in 86 primary studies that met the inclusion criteria. The list of relevant papers was continuously updated whenever a new paper was found.In the general search, four main keywords are used in the search: energy modelling, prediction, occupant's behaviour, and machine learning, taking into consideration their alternative spellings and synonyms to select any papers where they were found in the title, abstract and keywords. Papers that are not peered reviewed in the recognized publication are excluded, as well as papers that did not discuss the application of machine learning algorithms in building energy prediction. For more inclusion and specifics on the subject, a further search was made through the selected search engines using more subject-specific words. For more specific search, the search is broken down into the categorized occupant's actions and occupancy (windows opening, shades and blinds, plug loads or appliances, adjustments of thermostats, lighting), along with the keywords: energy modelling, prediction, and machine learning. This allowed us to narrow down our search to 86 publications specifically covering the use of selected machine

learning algorithm in predicting one or more occupant's behaviours and/or occupancy.

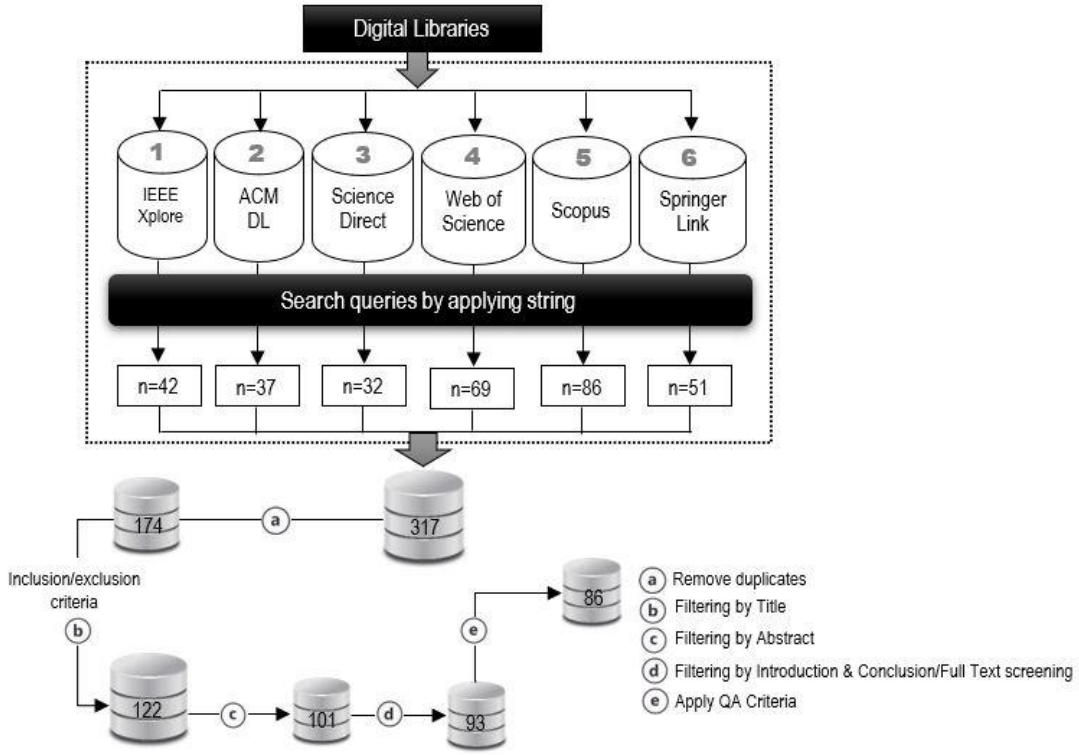


Figure 3-7 Search strategy

3.4.2 Inclusion Criteria

To ensure the relevance and usefulness of the selected studies, inclusion and exclusion criteria are applied in accordance with the research questions and search process. Studies meeting the following characteristics are included:

- written in the English language,
- centered around solo and ensemble models, methods, techniques, or approaches for energy consumption prediction,
- focused on energy consumption, prediction, forecasting, or calculation and published in peer-reviewed conferences or journals,
- utilization of machine learning algorithms in predicting energy consumption in residential buildings.

3.4.3 Exclusion Criteria

The studies included in this research were required to comply with certain characteristics. Specifically, studies had to be based on empirical research and consider the impact of occupant behavior on energy consumption. Additionally, only studies focused on machine learning techniques were included, and those not meeting these criteria were excluded.

3.4.4 Research Quality Valuation

To assess the quality of the primary studies, a customized version of the quality assessment checklist proposed by Kitchenham et al. (2007) was used. This checklist has been used in previous SLR studies such as those by Idri et al. (2016) and Wen et al. (2012). The customized checklist consists of 12 questions aimed at evaluating the quality, accuracy, dependability, and impact of the selected studies (table 3.4). Each question was scored on a three-point scale: Yes has 1 point, No has 0 points, and Partial has 0.5 points. Forming a max score of twelve. To be included in this study, a primary study had to achieve an acceptable quality score greater than the passing score of half the max score, any study below the average was excluded. A total of 7 papers failed to achieve the inclusion score of above 6. The quality scores of the remaining 62 primary studies are summarized in Table 3.5 and Figure 3.8.

Table 3-4 Quality assessment checklist

Question	Score
1. Are the objectives of the research clear?	1 y, 0 n, 0.5 p
2. Was the study designed to reach the stated aims?	1 y, 0 n, 0.5 p
3. Are the machine learning techniques used detailed and described?	1 y, 0 n, 0.5 p
4. Did it undertake variable measurements with reliability?	1 y, 0 n, 0.5 p

5. Did it include the data collection methodology?	1 y, 0 n, 0.5 p
6. In case the answer was yes, was it described comprehensively?	1 y, 0 n, 0.5 p
7. Was the study transparent?	1 y, 0 n, 0.5 p
8. Were evaluation metrics used to evaluate the mode?	1 y, 0 n, 0.5 p
9. Are the limitations discussed?	1 y, 0 n, 0.5 p
10. Does this relate to my research questions?	1 y, 0 n, 0.5 p
11. Did the model present any significant findings?	1 y, 0 n, 0.5 p
12. Are there any solid conclusions drawn from the results?	1 y, 0 n, 0.5 p

Table 3-5 Quality valuation of selected studies

ID	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	QA9	QA10	QA11	QA12	Score
1	1	1	1	0	0	1	1	1	1	1	1	1	9
2	0	1	0.5	0.5	0	0	0.5	1	1	0.5	0.5	0.5	5.5
3	1	1	1	0.5	0	1	1	1	1	0	1	1	9.5
4	1	1	1	0	1	1	1	0	1	1	1	0	8
5	0.5	1	1	1	0.5	1	0.5	1	0.5	0.5	0.5	1	8.5
6	1	1	0.5	1	0.5	1	1	1	1	0.5	1	1	9.5
7	1	1	1	1	0.5	1	0.5	1	1	1	0.5	1	9.5
8	1	1	0.5	0	1	1	1	1	0	0.5	1	1	8.5
9	0	0	1	1	0	0	0	1	1	0	0.5	1	5.5
10	1	1	1	0	0	1	1	1	1	1	1	1	9
11	0	1	0.5	0.5	0	0	0.5	1	1	0.5	0.5	0.5	5.5
12	1	1	1	0.5	0	1	1	1	1	0	1	1	9.5
13	0.5	1	1	0	1	1	1	0	1	1	1	0	8
14	0	1	1	1	0.5	1	0.5	1	0.5	0.5	0.5	1	8.5
15	1	1	0.5	1	0.5	1	1	1	1	0.5	1	1	9.5
16	1	1	1	1	0.5	1	0.5	1	1	1	0.5	1	9.5
17	1	1	0.5	0	1	1	1	1	0	0.5	1	1	8.5
18	1	0	0	0.5	1	0.5	1	1	1	0	1	0.5	7.5
19	1	1	1	1	1	1	1	1	0.5	0.5	1	0.5	10.5
20	1	0	0	0.5	1	1	1	1	0.5	0.5	0	0.5	7
21	1	0	0	1	1	1	1	1	0	1	1	1	9
22	1	0.5	0.5	0	0	1	1	1	0	0	0.5	1	6.5
23	1	1	1	1	1	0	0.5	0	1	1	0.5	1	9
24	1	0	0	0.5	0.5	0.5	1	0	1	1	1	0	6.5
25	1	1	1	1	0	0	1	1	0	0	0.5	0	6.5
26	1	1	1	1	1	0.5	0.5	1	1	0	0.5	0.5	9
27	0	1	1	1	0	0	1	0.5	0	1	1	1	7.5
28	1	1	1	1	1	0	1	1	0	1	1	0.5	9.5
29	1	0.5	1	1	1	0.5	1	1	0	0	1	0	8
30	1	1	1	1	1	1	1	1	0	0	1	1	10
31	1	1	1	0	0.5	0.5	1	1	0.5	0.5	0	0	7
32	0.5	0.5	1	1	1	1	0.5	0.5	0	0	1	1	8
33	1	1	1	0	0	1	1	1	0.5	1	0.5	1	9

32	0.5	0.5	1	1	1	1	0.5	0.5	0	0	1	1	8
33	1	1	1	0	0	1	1	1	0.5	1	0.5	1	9
34	1	0	1	1	1	1	1	0.5	0	0.5	0.5	1	8.5
35	1	1	0.5	1	1	0	1	0	0	0	1	0.5	7
36	1	1	1	1	1	0	1	1	0.5	0	1	0.5	9
37	0.5	1	1	1	0.5	1	0	1	0	0	0.5	1	7.5
38	1	1	0.5	0	1	1	0	0	0	0.5	1	1	7
39	1	1	1	0.5	0	1	1	0	1	1	0	0	7.5
40	1	0.5	1	0	1	1	0.5	1	0	1	0	1	8
41	0	1	1	1	0.5	0.5	0	1	0.5	0.5	0.5	0.5	7
42	1	1	1	0	1	1	1	0	1	1	0.5	0.5	9
43	0	1	1	1	1	1	1	1	0.5	0	1	0	8.5
44	1	1	0.5	1	1	1	0	1	0	1	1	1	9.5
45	1	0.5	0.5	1	0	0	1	0	1	1	0	1	7
46	1	1	0	0	1	1	0	0	0.5	0.5	1	0	6
47	0.5	0	1	0.5	1	0	0.5	0	1	1	0	0	6.5
48	1	0.5	1	0	0	1	1	1	0	1	0.5	0	7
49	1	0.5	0	1	1	1	1	1	1	0	0.5	1	9
50	0.5	1	1	1	0	0	1	1	0	1	0.5	0	7
51	1	0	0	1	1	1	1	0	0	0	1	0.5	6.5
52	1	1	1	1	0	1	1	1	0.5	1	1	1	10.5
53	1	1	0	0	1	0.5	1	1	0	1	1	1	8.5
54	1	0	1	1	1	1	1	1	0	0	1	1	9
55	0.5	1	1	1	0.5	1	0.5	0.5	0	0	0.5	0	6.5
56	1	1	0	0.5	1	0.5	1	1	1	1	1	1	10
57	1	0.5	1	0.5	1	1	1	1	0.5	0.5	0.5	1	9.5
58	1	0	0	0.5	1	1	1	1	1	0.5	0	1	8
59	0.5	0.5	1	1	0	0	1	1	0.5	1	0	1	7.5
60	0.5	1	0	1	1	0	1	0	1	0.5	0	1	7
61	0	1	0	0	1	1	1	0.5	1	1	1	1	8.5
62	1	0	0	0	1	1	0	0	1	1	1	0.5	6.5
Total	52.5	44	41.5	43	44	43.5	47.5	43	23.5	33.5	40.5	39.5	497
Avgerage	0.85	0.71	0.67	0.69	0.71	0.7	0.77	0.69	0.38	0.54	0.65	0.64	8.02

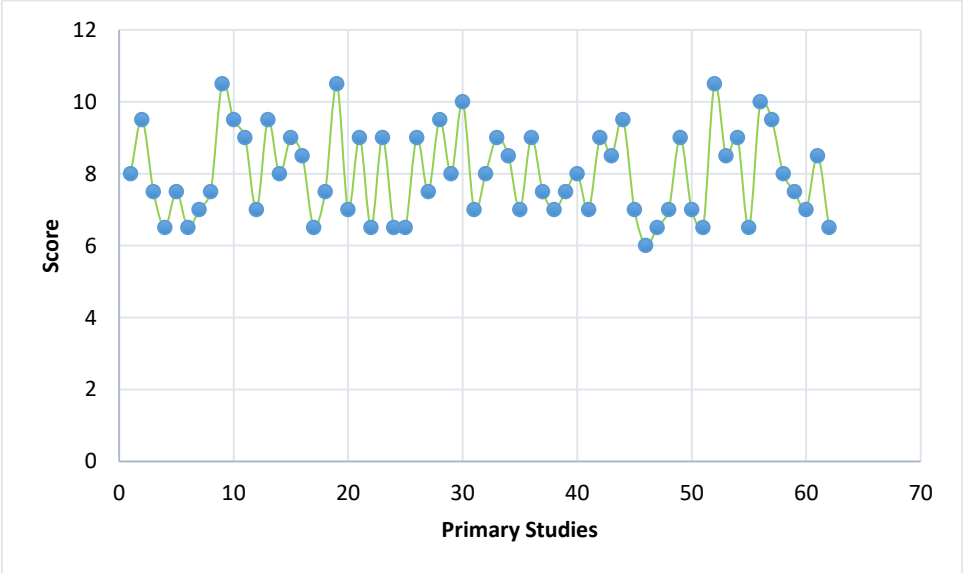


Figure 3-8 Quality score

3.5 PHASE-3: Model Development

The preliminary studies also defined the gaps in the current models resulting from the design of the new conceptual model for improving occupant's behavior-based energy consumption accuracy prediction using a combination of machine learning techniques to make an ensemble. Table 3.6

summarizes the detailed activities and deliverable of the model development phase of the operational framework.

Table 3-6 Phase 3 of operational framework

PHASE-3:		
Activities	resources	outputs
Phase 3.1. Design ensemble energy consumption prediction model	rationale and relationships and rational	model
Phase 3.2. Develop proposed ensemble energy consumption prediction model	numerical calculation	Ensemble energy consumption prediction model

Figure 3.9 shows the conceptual model with the input variables and occupant's characteristics. The predicted output is the total energy consumption, while the predictors are occupants, and occupant behaviour related parameters (listed below). The predictive model is built to predict the energy performance in dwellings based on the impact of occupant behavioral parameters:

i) Use of appliances, number of appliances
ii) Number of windows and doors
iii) Type of thermostats and occupants use of thermostats
iv) Type of air-conditioning, number of units, occupants use
v) Type and number of lightings, interaction with lighting units
vi) Occupant's characteristics (gender, age, number of occupants, education levels, income)
vii) Occupancy and presence during weekdays and weekends.

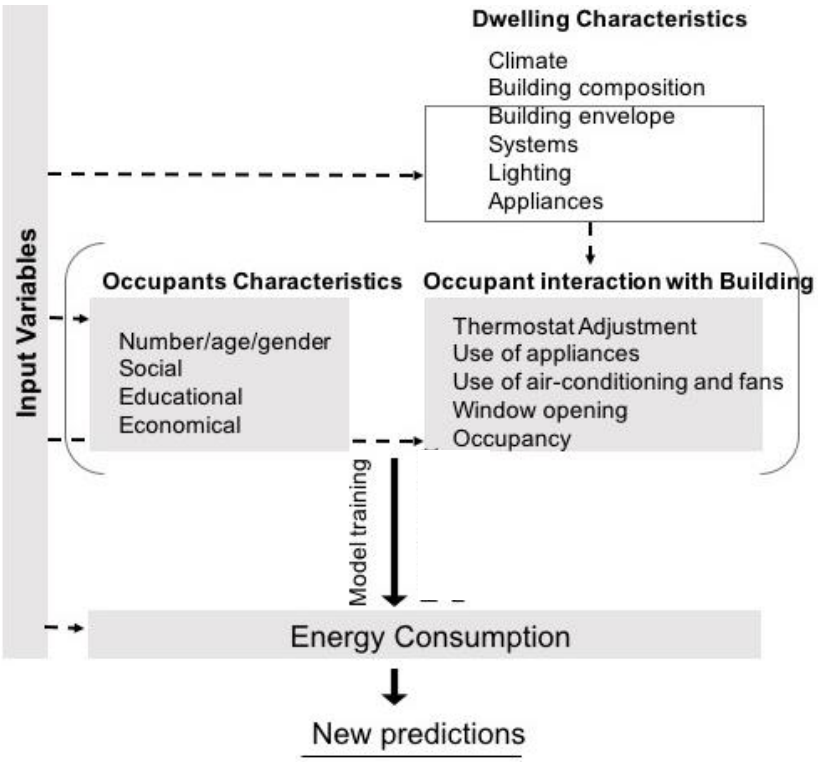


Figure 3-9 Conceptual model with occupants' behaviour

3.6 PHASE 4: Data Collection and Analysis

In this phase, the dataset from American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) global occupant behaviour database is collected and statistical analysis of the data is provided to generate results that will be evaluated and discussed. Table 3.7 provides a summary of the detailed activities and the expected outcome of this phase of the operational framework. The dataset is used to measure the solution by applying the collected data on the proposed ensemble model to compute predictive energy consumption. Data from several data sources reflecting the type and inclusiveness of data needed to develop the machine learning model. The dataset covers occupants related parameters, building related parameters as well as the energy consumption in buildings. The dataset types, plan and stages of evaluation are shown in Figure 3.10.

Table 3-7 Phase-4 of operational framework

PHASE-4: DATA COLLECTION AND ANALYSIS		
Activities	resources	outputs
data collection	ASHRAE global occupant behaviour database	Charts, graphs, Predictive energy consumption
analysis	Evaluation and simulation	results

The outcomes are analyzed by varius evaluation metrics. Those metrics, and the findings are presented in chapters 5 and 6.

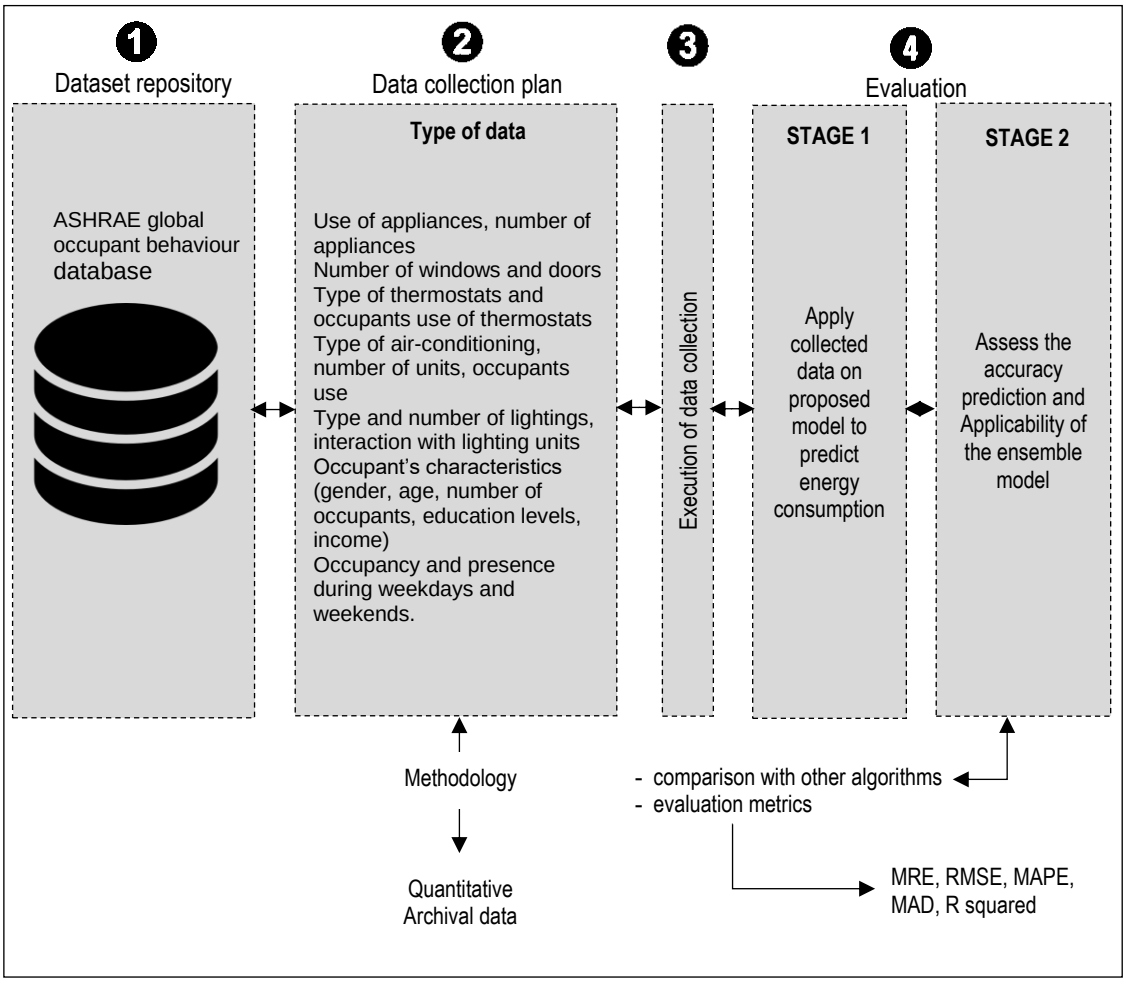


Figure 3-10 Structure of phase 4 and 5

3.7 PHASE 5: Evaluation

This phase addresses RQ3 and RQ4: During this phase, the proposed ensemble model is evaluated to determine its applicability. The evaluation process involves comparing the actual prediction with the prediction generated by the proposed model. Evaluation metrics are used to evaluate the model, and two evaluation aspects will be considered.

3.7.1 Evaluation 1

To address the RQ3, the first evaluation aspect is to assess the accuracy of prediction improvement of the ensemble model. The solo machine learning models are compared to the ensemble model. Evaluation metrics are employed to statistically analyze the results produced by the models.

In this evaluation, the following hypothesis is evaluated:

i. **Null Hypothesis (H0):** The energy consumption prediction accuracy of an ensemble prediction model is not better than the existing solo models.

ii. **Alternative Hypothesis (H1):** The energy consumption prediction accuracy of an ensemble prediction model is better than the existing solo models.

3.7.2 Evaluation 2

To answer RQ4, the proposed ensemble model's predictive reliability and applicability are evaluated using various commonly used evaluation metrics such as Magnitude of Relative Error (MRE), Mean Magnitude of Relative Error (MMRE), Mean Absolute Deviation (MAD), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).
In this evaluation aspect, the following hypothesis is evaluated:

i. **Null Hypothesis (H0):** The ensemble prediction model is not applicable to accurately predict the energy consumption of residential buildings.

ii. **Alternative Hypothesis (H1):** The ensemble prediction model is applicable to accurately predict the energy consumption of residential buildings.

The MRE is defined as the ratio of actual to predicted consumption. The MRE is calculated according to the below (3.1).

$$\text{MRE} = \frac{|\textbf{actual effort} - \textbf{estimated effort}|}{\textbf{actual effort}} \tag{3.1}$$

The MMRE value is calculated using the MRE values through the following equation(3.2).

$$\text{MMRE} = \frac{1}{n}\sum_{1}^{n} MRE_i \tag{3.2}$$

Where,

n = number of observations

The MMRE is utilised to detect the amount of predicted consumption to check the under-prediction or over-prediction attributes in assessment to the actual prediction. Because of its characteristic independent-of-units.

The mean squared error (MSE) is a common loss function used in introductory machine learning. To calculate the MSE, actual data is averaged across the entire dataset and the squared difference between the actual and estimated values is calculated. This is represented by the following equation(3.3).

$$\text{MSE} = \frac{1}{n}\sum_{1}^{n} e_i^2 \tag{3.3}$$

Where,

- n = number of observations and ‘e’ prediction error value

The root mean square, commonly referred to as the roots squared deviations, is one of the techniques more regularly that used evaluate the correctness of predictions. It displays the distance function among actual measurements and forecast. Compute the root-mean-square by calculating the residual (discrepancy among forecast as well as true) for each data point, as well as its norms, average, and squared. RMSE is often used in supervised machine learning algorithms since it demands and makes use of direct measures on every predicted data set. It is calculated by using the Equation (3.4).

$$\text{RMSE}= \frac{\sqrt{\sum_{i=1}^{N}\left|\left|y(i)-y^{i}\right|\right|2}}{N} \tag{3.4}$$

Where, ‘n’ is the number of observations.

Usually, range between all statistics point and the mean is known as an average relative difference of a dataset. It offers us a sense of how variables a dataset is. The mean absolute deviation can be calculated by using the Equation (3.5):

Step 1: Determine the Means

Step 2: Determine the appropriate ranges that every single item must be above the means.

Step 3: Adding those deviation equally

Step 4: Divides the total from the quantity of input values.

$$\text{MAD}=\sum\left|\frac{xi-x}{n}\right| \tag{3.5}$$

A means relative percentage errors could be used to assess an accurateness 's efficiency in algorithms. The MAPE is a lost functional which precisely identifies a model's mistake. The MAPE is calculated simply calculating the relative differences among the real and predicted numbers, then divided by the real values. These fractions for all variables are added to determine the means. More concisely, the formula for the MAPE is shown in Equation (3.6)

$$\text{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{A-F}{A}\right| \tag{3.6}$$

∑= urges combining the entire variables produced.

N= it representative sampling enough

a= it indeed the true worth

f= it an anticipated worth

3.7.3 Evaluation 3

Evaluation 3 aimed to compare the performance of the developed ensemble model in predicting the energy consumption of a building in kWh/m2/annum, to (1) the actual energy consumption of a real case study taken from Annex53 (IEA) and (2) the simulation results of the same case study conducted on DesignBuilder by using existing inputs, schedules, and assumptions of missing occupant behavioral parameters, and predicting energy consumption using EnergyPlus.

3.8 Summary

The chapter provided an indepth outlook of research methodology.. The engineering method is used as a research methodology. The engineering method is performed in four steps which are (1) observe the existing solutions; (2) propose a better solution; (3) develop the proposed solution; and (4) measure and analyze. This section provides an elaboration on the research design, which includes the research structure, decision-making process, and its three main phases. It also highlights the operational framework, which consists of five phases and provides a detailed explanation of each of these phases. The proposed ensemble model is designed and clarified in each base model with its functionality to create an ensemble.

The archival data from American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) global occupant behavior database is

used as the data collection method. The case study is adopted from annex53 (IEA) and simulated to compare results. The research process involved a structured approach to data collection, evaluation, and analysis. Through this process, the accuracy and applicability of the proposed ensemble energy consumption prediction model were determined.

AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE:
CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

ENSEMBLE MODEL ARCHITECTURE AND ALGORITHMIC DESIGN

Chapter 4

Chapter 4

ENSEMBLE MODEL ARCHITECTURE AND ALGORITHMIC DESIGN

4.1 Introduction

This chapter presents the development and overview of the proposed ensemble model to improve prediction accuracy of energy consumption of residential buildings by integrating all the factors influencing occupant's behaviours. The discussion in this chapter is structured to four-layer units used in this model on conceptual and mathematical relationship justification of the base models such as 1) Lasso regression, 2) Ridge regression 3) Random forest regressor 4) Gradient Boost. Lastly a summary is presented to conclude the chapter.

4.2 Lasso regression

Lasso regression is a linear regression algorithm using L1 regularization so that the model can have sparse coefficients. It penalizes the large coefficients, and drives the model to rely only on the most important predictors for making predictions. Lasso regression is particularly employed when the dataset is highly dimensional with many potential predictors but only a few are truly important. This aligns with our research aim of predicting energy use based on the highly dimensional number of predictors. Lasso is select the most important predictors for the model inorder to provide more accurate predictions.

The equation for Lasso regression is represented as follows:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp + \varepsilon$$

where;

y represents the response variable,

β0 represents the y-intercept or constant,

β1 to βp represents the regression coefficients for the predictor variables x1 to xp,

and ε represents the error term.

Inaddition, Lasso regression uses regularization, which is represented by λ. The equation with the regularization term is:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp + \varepsilon + \lambda * \sum|\beta i|$$

where;

λ is the regularization parameter,

and |βi| represents the absolute value of the regression coefficient for the ith predictor variable.

The L1 regularization aids in reducing the variance of the model and improving the predictive accuracy.

The Lasso regression process can be divided into the following parts:

Part 1: similar to any algorithm, data need to be processed where missing values are removed, data is scaled and categorical variables are encoded if they are available.

Part 2: feature selection, which includes the penalty term that shrinks the least important coefficients value to zero, focusing only on the most important features or predictors.

Part 3: After features are selected, the data is split into training and testing datasets.The lasso model is then trained on the training dataset, in which the model will learn the selected features or predictors coefficients and the optimal regulization parameters.

Part 4: After training the model using the training dataset, the model performance is evaluated on the testing set for validation. This means the model is performing on new not trained on data.

Part 5: The model performance can be assessed by evaluation metrics, and if the results are not satisfactory, the regularization parameters can be tuned and the model can be evaluated again until the results are satisfactory.

Part 7: *once the model is satisfactory, the model can be used on new data to provide prediction.*

Figure 4.1 provides a flow chart of the Lasso regression model architecture.

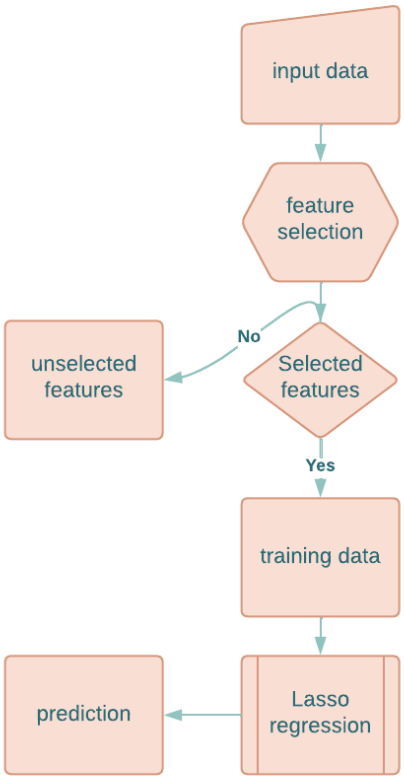


Figure 4-1 Lasso Regression

4.3 Ridge regression

Ridge regression, similar to lasso, is another linear regression technique. However it uses the L2 regularization so the model can avoid being overfitted. Similar to Lasso, λ is the regularization term. The Ridge regression equation with the regularization term is as follows:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp + \varepsilon + \lambda * \sum(\beta i^2)$$

where;

λ is the regularization term,

and βi^2 represents the squared value of the regression coefficient for the ith predictor variable.

The Ridge regression algorithm aims to reduce the sum of squared residuals. This in return, helps in reducing the variance of the prediction model and improving overall accuracy. Ridge also deals with high dimensional data with various predictors, such as our case.

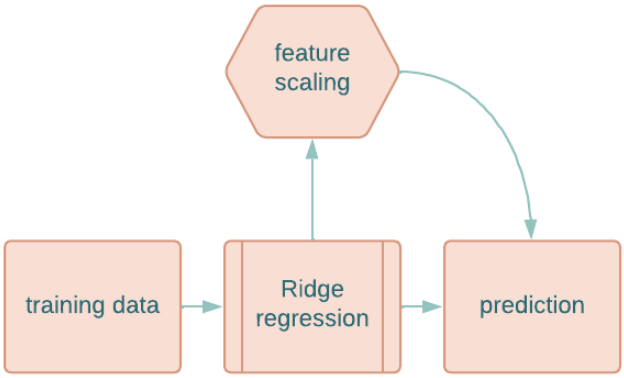


Figure 4-2 Ridge regression

The Ridge regression components are presented in figure 4.2 and can be summarized into the following parts:

Part 1: cleaning the data and processing, splitting into training and testing datasets.

Part 2: feature scaling, where the data is standardized to zero means and unit variance.

Part 3: After features are scaled, the model is then defined including its hyperparameters.

Part 4: the model is then fitted by finding its optimal hyperparameters.

Part 5: The model performance is then evaluated using evaluation metrics.

Part 6: Hyperparameters are tuned to improve the model performance if needed.

Part 7: *The optimal model is chosen to perform predictions.*

4.4 Random Forest

Random Forest is an ensemble learning algorithm that structures multiple decision trees at training time and yields the class of the individual trees. Random Forest is known to handle complex datasets and recognize the most important features for the predictions. RF is also able to handle missing data while at the same time maintaining high prediction accuracy. RF constructs multiple decision trees by using bootstrapped samples of the dataset and then randomly selecting subsets of features to build each tree. This is used to

aggregate the predictions from the n number of trees to make a final prediction. The Random Forest algorithm can be represented as follows:

Part 1: cleaning the data and processing..

Part 2: Random selecting features subsets from the dataset.

Part 3: After features subsets are selected, decision trees are created using bootstrapped data sample from the feature subsets.

Part 4: the second and third part of this process is reiterated to create a forest of decision trees.

Part 5: The prediction of each tree is aggregated to form the final prediction. The prediction is based on the average of all trees.

Figure 4.3 draws the process and architecture of Random Forest.

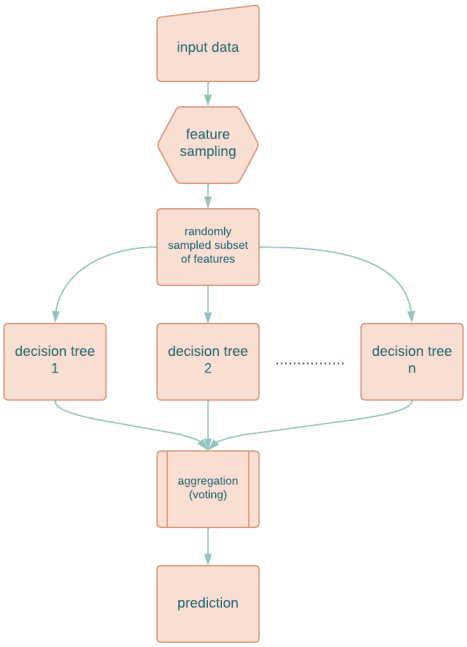


Figure 4-3 Random Forest

4.5 Gradient Boosting Regression

Gradient Boosting is another machine learning technique that can handle both; classification and regression. It combines multiple models which are weak and learn everytime to produce stronger model. GBR adjusts the weights of the parameter points based on residual errors of the models. The errors are then

calculated between the predicted values and the actual values. These residual errors are then used to train a new model that is added to the ensemble. The iteration process is repeated n times until it reaches the final model which is satisfactory. Gradient Boosting exhibits many hyperparameters, which can be tuned to improve its performance, such as the learning rate, the number of trees in the ensemble, and the maximum depth of the decision trees. Figure 4.4 describes the process of gradient boosting.

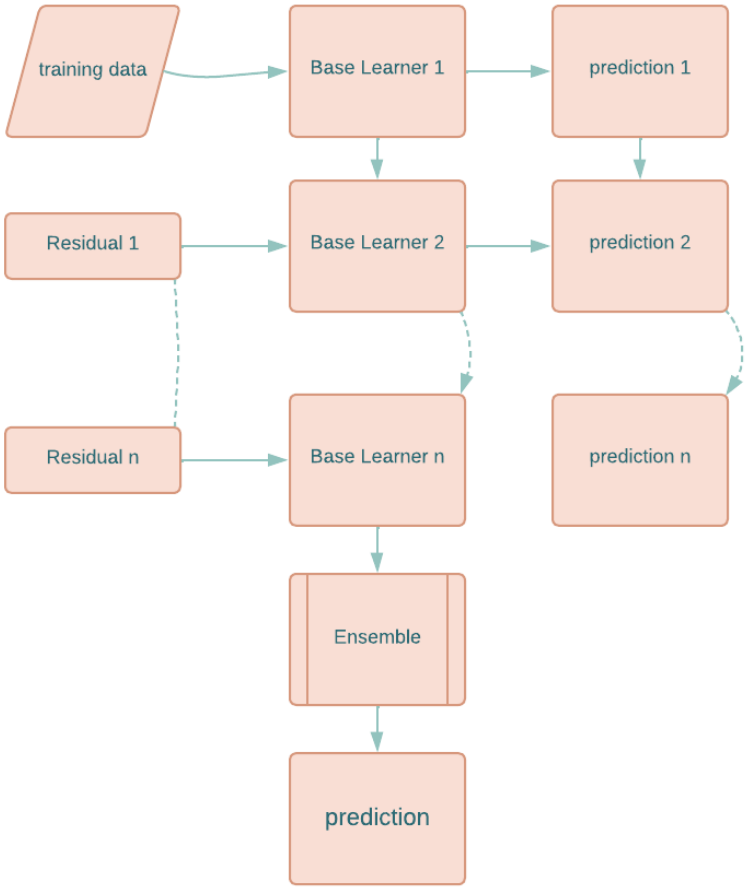


Figure 4-4 Gradient Boosting

The steps followed by gradient boosting algorithm can be summarized by the following parts:

Part 1: cleaning the data and processing..

Part 2: Make n1 prediciton through fitting a simple decision tree to the model.

Part 3: calculate the residual error of the n1 prediction.

Part 4: Boosting, which means fiting a new model n2 to the residuals of the first model n1. This leads to the model learning and improving predictions.

Part 5: iterate this process n times to improve predictions.

Part 6: combination of all models to obtain the final model prediction.

4.6 Proposed Model

The preliminary studies also defined the gaps in the current models resulting from the design of the new proposed model for improving occupant's behaviour-based energy consumption accuracy prediction using a combination of machine learning techniques to make an ensemble using Lasso, Ridge, RF, and GBR.

Over the past few decades, lots of research have been conducted on various types of energy consumption prediction techniques and lot of models have been proposed to achieve high accuracy. To overcome the drawbacks and combine the strengths of energy consumption prediction techniques, a new technique called ensemble prediction has been explored. It consists of combining more than one technique to predict the energy consumption by means of a combination rules. Based on the usage of ensemble or combinations of methods, this methodology is now being applied to predict tasks in data mining. According to data mining studies, ensemble methods produce more accurate results than single methods. This has inspired the researchers to use ensemble methods in various fields. The basic idea behind using ensemble prediction is that each single technique has its merits and demerits, we can minimize the limitations by integrating techniques via ensemble prediction, which may lead to more accurate prediction. Table 4.1 explores these merits and demerits of our solo algorithms and the ensemble model. The table is derived based on evidence from the extensive literature review in chapter 2 and common understanding of each algorithm.

Table 4.1 Solo ML versus Ensemble merits and demirts

Feature / Model	Ridge Regression	Lasso Regression	Random Forest	Gradient Boosting	Ensemble Model
High-dimensional datasets		X	X	X	X
Nonlinear relationships			X	X	X
Sparse data		X			X
Outliers	X	X	X	X	X
Feature selection		X	X	X	X
Interactions between features			X	X	X
Missing data	X	X	X	X	X
Heterogeneous data			X	X	X
Time-series data	X	X		X	X
Categorical data		X	X	X	X

Ensemble techniques use combination rules such as mean, median, Inverse Rank Weighted Mean, etc. to create an ensemble. These methods can be classified into two categories:

i) Homogeneous: used to refer to an ensemble that consolidates one base model with no less than two distinct combinations of one ensemble learning.

ii) Heterogeneous: used to refer to an ensemble of two or more different base models.

Researchers have piloted various empirical studies to assess ensemble energy consumption techniques. Some of these studies were dedicated to dealing only with homogeneous ensembles, heterogeneous ensembles, or both types of techniques. Each base technique can compensate for prediction errors made by other base methods. The ensemble energy consumption prediction process is shown in Figure 4.5.

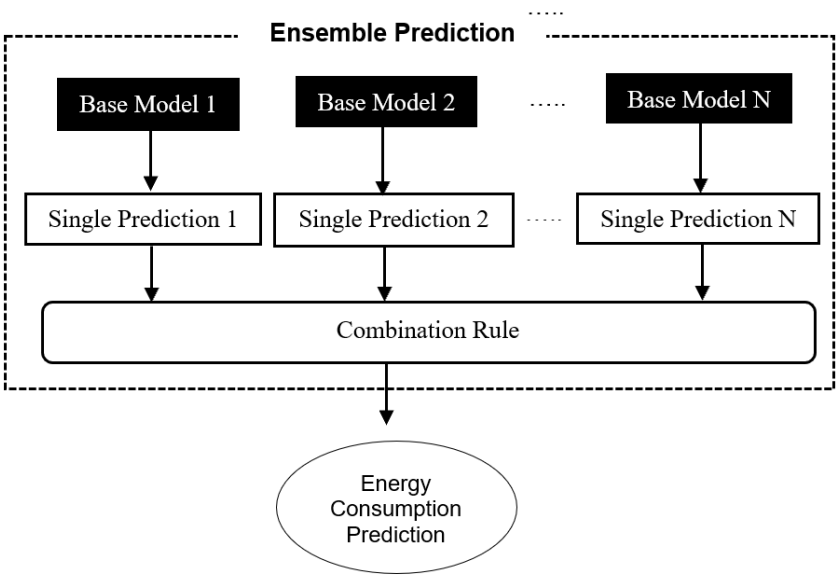


Figure 4-5 Ensemble energy consumption prediction process

Figure 4.6 shows the ensemble process of our ensembles ridge, lasso, RF, and GBR algorithms. The predicted output is the total energy consumption, while the predictors are building, occupants, and occupant behavior related parameters. The predictive model is built to predict the energy performance in residential buildings based on the impact of occupant behavioral parameters.

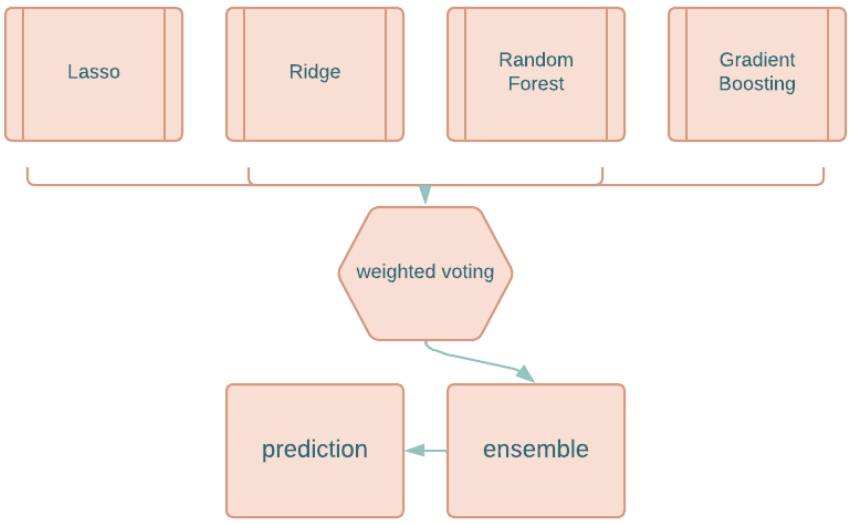


Figure 4-6 ensemble process

The model takes in the input parameters including the use of appliances and the number of appliances, the number of windows and doors, the type of thermostats and occupants' use of thermostats, the type of air conditioning, the number of units, the occupants' use, the type and number of lightings, the interaction with lighting units, the occupant's behavior, and occupancy and presence during weekdays and weekends. A full breakdown of the dataset description is presented in chapter 5. The input data then ensues to each individual model, which are the chosen models; Lasso Regressor, Ridge Regressor, Random Forest Regressor, and Gradient Boosting. The models then produce their respective predictions based on the steps detailed in this chapter above. The predictions from each model are then combined using the Weighted Voting Regressor. The weighted voting regressor assigns weights to all the models according to their individual performance. The weighted predictions are then combined to generate the final Energy Use Prediction. Since energy use prediction is calculated based on the four base models, an ensemble model is then formed.

The breakdown of our model in detail is presented in the following parts:

Part 1: Cleaning the data and processing. This includes: dealing with missing values, creating dummy variables, splitting data into numerical and categorical data, dealing with skewness, and removing outliers.

Part 2: The input data is pre-processed using the 'StandardScaler' function to transform the data into a standard normal distribution.

Part 3: The preprocessed data is then fed into the four different regression models: Lasso, Ridge, Gradient Boosting, and Random Forest. Each model has their own defined sets of hyperparameters. Lasso and Ridge models use regulation techniques to select important features and prevent overfitting. Random Forest and Gradient Boosting combine multiple models to improve overall prediction accuracy and avoid underfitting.

Part 4: After the four models are being fitted, a weighted voting regressor model combines their predictions. The model then provides the final prediction based on the weighted average.

Part 5: The output of the ensemble model is an energy use prediction.

4.7 Summary

The proposed ensemble machine learning approach has been presented in this chapter which gives the answer to RQ2.

"How to develop an occupancy behavior-based ensemble machine learning model to improve energy consumption accuracy prediction of residential buildings using Lasso regression, Ridge regression, Random Forest, and Gradient boosting?"

A novel ensemble machine learning model to improve the prediction accuracy of energy consumption of residential buildings that combines these algorithms is proposed. The prediction results of the four algorithms are ensembled weighted voted regressor. Finally, the energy consumption prediction is calculated based on four base models, hence an ensemble model is developed.

AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE:
CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

MODEL DEVELOPMENT AND EVALUATION

Chapter 5

Chapter 5

MODEL DEVELOPMENT AND EVALUATION

5.1 Introduction

This chapter provides a comprehensive overview of the entire model development process. Firstly, it covers the collection of datasets, including their source and a detailed description of the datasets, highlighting their features, size, and structure, as well as any potential issues or limitations.

Next, the chapter delves into the exploratory data analysis process. Exploratory data analysis is performed to examine the dataset including but not limited to, structure, patterns, and trends. This allows a deeper understanding of the data, especially in terms of quantifying the occupant behavior's impact on energy performance. This leads to making informed decisions about the modeling process.

The data pre-processing stage is also discussed in detail. This includes techniques such as addressing skewness, handling missing values, and identifying and addressing outliers. Pre-processing is a crucial step to ensure that the data is in a suitable form for carrying out the modeling stage.

The chapter then moves on to the process of evaluating algorithms. This involves training and comparing several algorithms and assessing their performance by means of evaluation metrics.

The next step is building an ensemble model using the selected models. The models are selected based on their evaluation scores and their merits when ensembled (refer to table 4.1). The models are combined to form an ensemble with improved accuracy in the predictions.

Finally, the chapter concludes with the validation and evaluation of the model. This involves assessing the model's performance on a separate validation dataset to ensure that it is robust and generalizable. The chapter also covers techniques for interpreting the results and identifying areas for improvement in the model.

5.2 Dataset Description

5.2.1. Data Acquisition

In building energy use data acquisition, especially in the case of developing machine learning models, acquiring a comprehensive dataset covering all factors that affect energy performance and occupant behavior can be challenging (Seyedzadeh et al., 2020; Zhang et al., 2022). The process of collecting big datasets is a dissuading task since it requires a significant number of resources and time. Moreover, the dataset collections should cover a range of years, buildings, and locations collected to ensure that the model has sufficient data to provide accurate results. In addition, occupant behavior data is usually collected through IoT sensors which can hinder privacy, especially in a residential setting (Jiang et al., 2021; Sayed et al., 2022). Hence, conducting a primary data collection process can be expensive and time-consuming, which is impractical for most solo researchers. Therefore, researchers often acquire publicly available datasets that have been collected and validated by other institutions to overcome these limitations. For my research objectives, various datasets were investigated in terms of suitability for model development. Table 5.1 summarizes available datasets considered based on size and input parameters.

Table 5.1 Occupant behavior and energy performance datasets

Dataset/ database Name	Description	Source	Sample Size
Commercial Building Energy Consumption Survey (CBECS)	National survey of energy use and related building characteristics	Energy Information Administration (EIA, 2021)	~5,600 buildings
ASHRAE Global Thermal Comfort Database II	Thermal comfort survey data collected from around the world	American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE, 2021)	~21,000 responses
UCI Appliances Energy Prediction	Energy use of home appliances	UCI Machine Learning Repository (Repository., 2020)	20,000 observations
ASHRAE Global Occupant Behavior Database	Occupant behavior survey data collected from around the world	American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE, 2018)	~10,000 responses

Residential Energy Consumption Survey (RECS)	National survey of energy use and related building characteristics	Energy Information Administration (EIA, 2021)	~5,600 household

Asessing suitable database was undertaken based on the following criteria:

1. Size of database, level of completeness and details.
2. Relevance of data related to occupant behavior parameters as well as occupants.
3. Validity and reliability of the data collection methods in the database
4. Diversity of case studies in the dataset
5. Accessibility of the dataset.
6. Ease of understanding of data.

5.2.2. Data Description

The dataset (ASHRAE, 2018) reflected the type and inclusiveness of data needed to develop the ensemble machine learning. The data covers occupants related parameters, building related parameters as well as the energy consumption in buildings as detailed in table 5.2 and table 5.3.

Table 5.2 Building related parameters

DOEID	Unique identifier for each respondent
TOTROOMS	Total number of rooms in the housing unit, excluding bathrooms
DOOR1SUM	Number of sliding glass doors
WINDOWS	Number of windows
NUMFRIG	Number of refrigerators used
NUMFREEZ	Number of separate freezers used
STOVEN	Number of stoves
STOVE	Number of separate cooktops
OVEN	Number of separate ovens
MICRO	Microwave oven used
DISHWASH	Have dishwasher
CWASHER	Have clothes washer in home
DRYER	Have clothes dryer in home
DESKTOP	Number of desktop computers
NUMLAPTOP	Number of laptop computers
NUMTABLET	Number of tablet computers or e-readers
INTERNET	Internet access at home
EQUIPM	Main space heating equipment type
THERMAIN	Any thermostats
AIRCOND	Air conditioning equipment used
COOLTYPE	Type of air conditioning equipment used
THERMAINAC	Thermostat for central air conditioner

PROTHERMAC	Programmable thermostat for central air conditioner
LGTINNUM	Number of light bulbs installed inside the home
SMARTTHERM	Smart thermostat
total area sqf	Total square footage (used for publication)
KWH	Total site electricity usage, in kilowatthours, 2015

Table 5.3 occupant related parameters

OVENUSE	Frequency of use of oven part of stove
NUMMEAL	Frequency hot meals are cooked
DWASHUSE	Frequency of dishwasher use
WASHLOAD	Frequency of clothes washer use
DRYRUSE	Frequency of clothes dryer use
TVCOLOR	Number of televisions used
TVONWD1	Most-used TV usage on weekdays
TVONWE1	Most-used TV usage on weekends
HEATHOME	Space heating used
EQUIPMUSE	Main heating equipment household behavior
NUMWHOLEFAN	Number of whole house fans used
FUELH2O	Fuel used by main water heater
LGTIN4	Number of inside light bulbs turned on at least 4 hours a day
SMARTMETER	Home has an electricity smart meter
USECENAC	Central air conditioner household behavior
NHSLDMEM	Number of household members
NUMADULT	Number of household members age 18 or older
NUMCHILD	Number of household members age 17 or younger
ATHOME	Number of weekdays someone is at home
PROTHERMAC	Programmable thermostat for central air conditioner
PROTHERM	Programmable main thermostat
INTDATA	Household has access to smart meter interval data

5.2.3. Data Processing

Dataset exploration and analysis forms the first step to model the relationship between energy use and the input parameters. The goal of this step is to analyse the dataset and identify the factors that affect energy use. This eventually leads to the development of the model and accurate energy use prediction.

The data processing, analysis, and model development were performed in a Jupyter notebook using Python, as Python provides numerous libraries for machine learning, data analysis, and processing.

a. Data examination:

Data examination is a crucial step in the data analysis and modelling process. It involves cleaning and transforming raw data into a format that is suitable for handling.

The dataset consists of 8741 columns which are the studied buildings sorted by different IDs, and 49 columns which are 22 related to occupants and occupants behavior, 26 related to building characteristics and the energy use in kwh. Figure 5.1 presents a snippet of this dataset.

	TOTROOMS	DOOR1SUM	NUMFRIG	NUMFREEZ	STOVEN	OVENUSE	STOVE	OVEN	MICRO	NUMMEAL	...
0	7	2	1	1	1	0.0	0	0	1	3	...
1	4	0	1	1	1	5.0	0	0	1	3	...
2	9	1	2	0	2	14.0	0	0	2	2	...
3	7	4	2	1	1	2.0	0	0	1	6	...
4	6	1	2	0	1	0.0	1	1	1	0	...

Figure 5-1 snippet of dataset

b. Replacing missing or invalid values:

The script checks for missing values in the dataset, and prints the number of columns with missing values for each type of column. : The script fills missing values in specific columns using imputation methods which replaces missing values with mean values.

c. Defining categorical and numerical columns:

The script then divides the dataset columns into categorical and numerical columns, and prints the number and names of each type of column.

d. Creation of indicator variables

Dummy variables, also known as indicator variables, are created to convert categorical variables into a numerical format in such a way it can be processed for machine learning application. This presents the categorical data in a numerical format.

e. Handling data skewness

The skewness of variables is checked and addressed. Skewness refers to the asymmetry in the distribution of the data. Figure 5.2 shows a sample of the skewed data. When the distribution is not normally distributed, a negative

impact can be achieved on certain algorithms. Log transformation is used as a common technique for handling skewed data.

The results in figure 5.3 show that the skewness of each variable has been reduced after applying the log transformation. This indicates that the distributions of the transformed variables are closer to a normal distribution, which can improve the performance of machine learning algorithms.

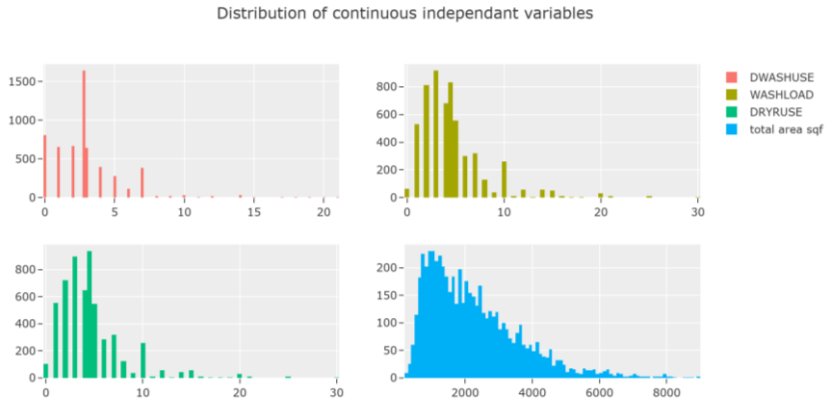


Figure 5-2 Data skewness

```
DWASHUSE skew before transforamtion: 2.123548
DWASHUSE skew after log transforamtion: -0.507073
WASHLOAD skew before transforamtion: 2.298941
WASHLOAD skew after log transforamtion: -0.004457
DRYRUSE skew before transforamtion: 2.249950
DRYRUSE skew after log transforamtion: -0.163159
```

Figure 5-3 handling data skewness

f. Handling outliers

outliers in the data are detected and removed using z-scores. Z-scores are a measure of how far away a data point is from the mean of the dataset, in terms of standard deviations. A z-score greater than 5 or less than -5 indicates that the data point is more than 5 standard deviations away from the mean, which is a very extreme value and is likely an outlier.

5.2.4. Data analysis

The distribution of values in the dataset is visualised for identifying patterns or issues in the data. Heatmap is developed as well to visualise the relationship between parameters. The code appendix A provides extensive insights on the

analysis. Figure 5.4 shows the histogram of energy consumption distribution in the dataset.

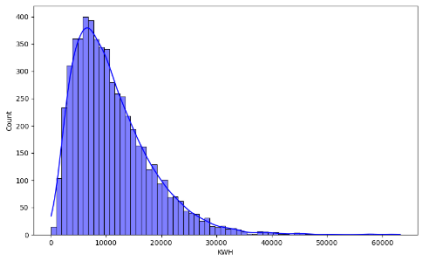


Figure 5-4 Histogram of energy use distribution

To understand the impact of occupant behavior on the energy use, a simple linear regression model is trained on the whole variables and then the impact of occupant behavioral variables is calculated. Firstly the data is split into features x and target y and the absolute value of the coefficient is determined and sorted according to impact (Fig 5.5).

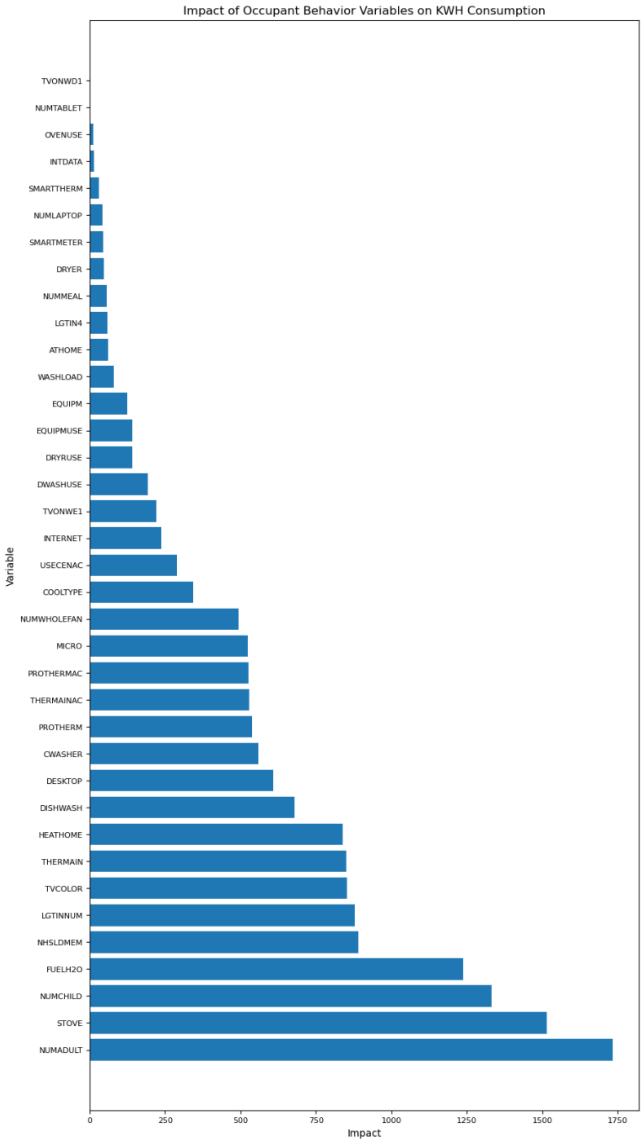


Figure 5-5 weighted impact of occupant behavior on energy performance

5.3 Modeling Phase

5.3.1. Algorithms selection

Following the data processing, The dataset is split into training and testing sets, which allowed us to run experiments and identify the optimal algorithms for our ensemble model A variety of models are fitted. The algorithms selection was based on the literature review of machine learning algorithms of chapter 2. Figure 5.6 shows the trained algorithms and Figures 5.7 and 5.8 provide box plots of the cross-validation results for the algorithms based scoring metric.

```
Training:  Lasso
Training:  Ridge
Training:  HuberRegressor
Training:  AdaBoostRegressor
Training:  RandomForestRegressor
Training:  ExtraTreesRegressor
Training:  GradientBoostingRegressor
Training:  SVR
Training:  DecisionTreeRegressor
Training:  KNeighborsRegressor
Training:  KernelRidge
Training:  XGBRegressor
Training:  LGBMRegressor
```

Figure 5-6 algorithm training

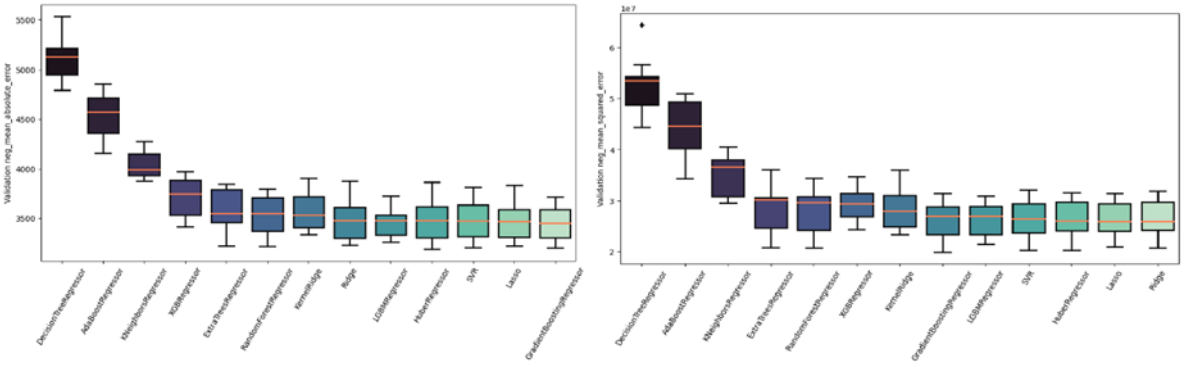


Figure 5-7 MAE and MSE evaluation metrics

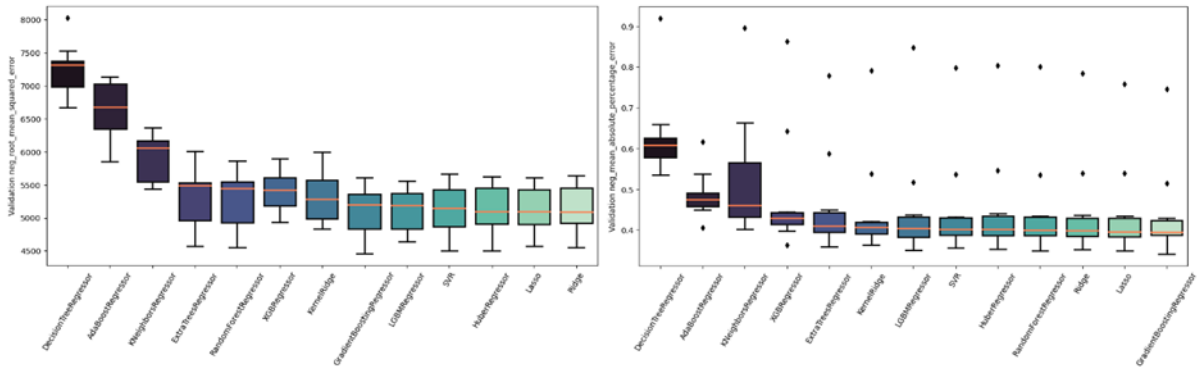


Figure 5-8 RMSE and MAPE evaluation metric

5.3.2. Choice of ensemble

Based on the results of the evaluation metric and the merits of each algorithm the choice of model was an ensemble of Lasso, ridge, Rf and gradient boosting. The algorithms performed well in terms of both mean absolute error and mean squared error and have also achieved relatively lower root mean squared error and higher coefficient of determination (R-squared) values. Therefore, these models might be good candidates for further analysis and selection.

Ensemble model improves prediction accuracy and robustness by combining the chosen models. Moreover, the ensemble voting employs weighting voting to inform the prediction values.

5.3.3. Ensemble Model building

The data is split into training and testing datasets with the features (X) and target variable (y) representing the input variables and the output energy use in kWh from the dataset. The data is then split into training and testing sets using a 70-30 split ratio. Four different pipelines are developed, each using a different regression algorithm: Ridge regression, Lasso regression, Gradient Boosting regression, and Random Forest regression. For each pipeline, the code defines a set of hyperparameters to be tested using a grid search with 5-fold cross-validation. After fitting each pipeline to the training data, the code generates predictions on the testing data (figure 5.9). The below metrics are used to evaluate the accuracy of our model in comparison to the individual algorithms.

```python
# Create a pipeline with Ridge regression
ridge_pipe = Pipeline([('scaler', StandardScaler()), ('ridge', Ridge())])
ridge_params = {'ridge__alpha': [0.001, 0.01, 0.1, 1, 10]}
ridge_grid = GridSearchCV(ridge_pipe, ridge_params, cv=5)
ridge_grid.fit(X_train, y_train)
ridge_best = ridge_grid.best_estimator_
ridge_best.fit(X_train, y_train)
ridge_pred = ridge_best.predict(X_test)
ridge_rmse = np.sqrt(mean_squared_error(y_test, ridge_pred))

# Create a pipeline with Lasso regression
lasso_pipe = Pipeline([('scaler', StandardScaler()), ('lasso', Lasso())])
lasso_params = {'lasso__alpha': [0.001,0.01, 0.1, 1, 10]}
lasso_grid = GridSearchCV(lasso_pipe, lasso_params, cv=5)
lasso_grid.fit(X_train, y_train)
lasso_best = lasso_grid.best_estimator_
lasso_best.fit(X_train, y_train)
lasso_pred = lasso_best.predict(X_test)
lasso_rmse = np.sqrt(mean_squared_error(y_test, lasso_pred))

# Create a pipeline with Gradient Boosting regression
gb_pipe = Pipeline([('scaler', StandardScaler()), ('gb', GradientBoostingRegressor())])
gb_params = {'gb__n_estimators': [50, 100, 200], 'gb__learning_rate': [0.1, 0.01, 0.001]}
gb_grid = GridSearchCV(gb_pipe, gb_params, cv=5)
gb_grid.fit(X_train, y_train)
gb_best = gb_grid.best_estimator_
gb_best.fit(X_train, y_train)
gb_pred = gb_best.predict(X_test)
gb_rmse = np.sqrt(mean_squared_error(y_test, gb_pred))

# Create a pipeline with Random Forest regression
rf_pipe = Pipeline([('scaler', StandardScaler()), ('rf', RandomForestRegressor())])
rf_params = {'rf__n_estimators': [50, 100, 200], 'rf__max_depth': [5, 10, None]}
rf_grid = GridSearchCV(rf_pipe, rf_params, cv=5)
rf_grid.fit(X_train, y_train)
rf_best = rf_grid.best_estimator_
rf_best.fit(X_train, y_train)
rf_pred = rf_best.predict(X_test)
rf_rmse = np.sqrt(mean_squared_error(y_test, rf_pred))
```

Figure 5-9 snippet of pipelines

The final prediction is then obtained by aggregating the predictions of all the models, using a voting scheme (Fig 5.10). There are several types of voting schemes that can be used in ensemble voting, including:

Majority Voting: In this scheme, the final prediction is the one that is predicted by the majority of the individual models. This is the most commonly used voting scheme in ensemble voting.

Weighted Voting: In this scheme, each individual model is assigned a weight, and the final prediction is obtained by taking a weighted average of the predictions of all the models.

The weighted voting is then performed to combine the 4 pipelines and build the ensemble mode.

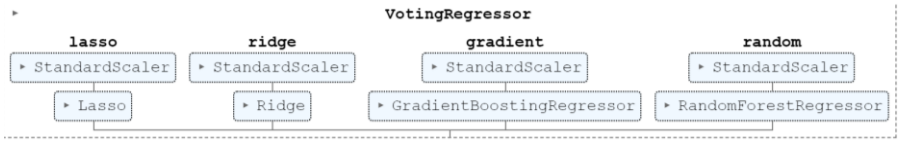


Figure 5-10 ensemble model pipeline

5.3.4. Model Evaluation

The table 5.4 shows the performance metrics of the different regression algorithms for the energy use prediction model. The metrics used for evaluation are R-squared, RMSE, recall, MAE, and MAPE. The algorithms

evaluated are Ridge regression, Lasso regression, Gradient Boosting regression, Random Forest regression, and an ensemble method. The ensemble method shows the best performance with an R-squared of 0.74, RMSE of 3475.88, recall of 1.0, MAE of 3061.36, and MAPE of 0.32. On the other hand, The solo algorithms show lower performance in comparison to the ensemble.

Table 5.4 evaluation metrics for solo and ensemble models

Model	R-squared	RMSE	Recall	MAE	MAPE
Ridge	0.531048	4524.985266	0.991587	3329.832761	0.577304
Lasso	0.532457	4518.183342	0.992188	3322.499182	0.674544
Gradient Boosting	0.624321	4557.326378	1.000000	3308.138354	0.558364
Random Forest	0.509503	4627.763719	1.000000	3401.046959	0.462632
Ensemble	0.741171	3475.882982	1.000000	3061.362711	0.322259

5.3.5. Model Results

Goodness of fit is essential for evaluating linear regression models and is determined based on linearity, homoscedasticity, and homogeneity criteria (Fig 5.11).

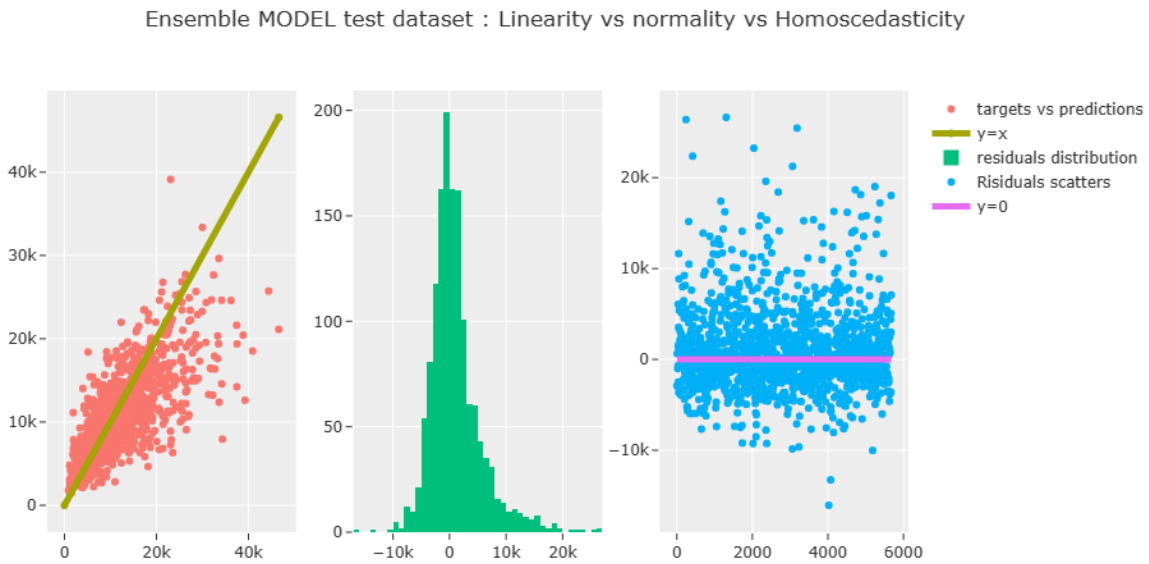


Figure 5-11 linearity vs normality vs Homoscedasticity

Our model exhibited excellent goodness of fit, as demonstrated in the figure 5.12, which shows the actual versus predicted values for the test set.

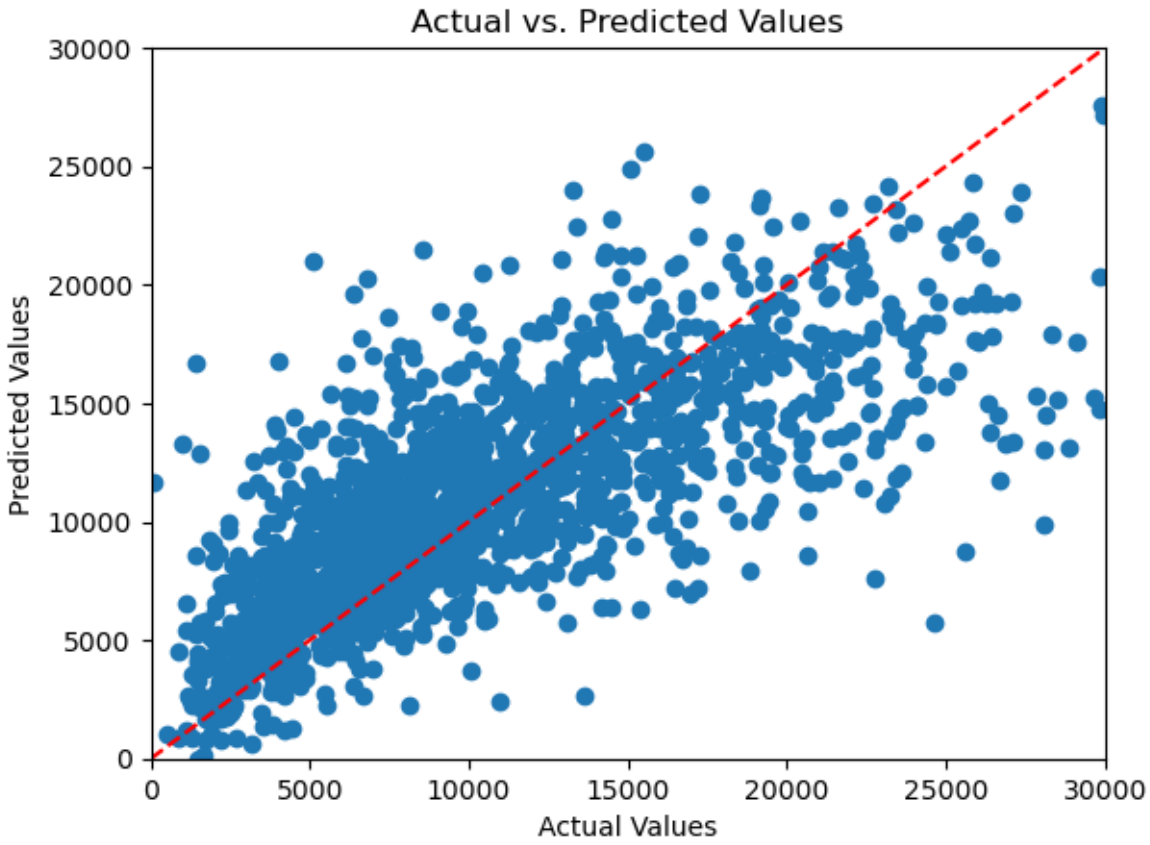


Figure 5-12 actual vs predicted

Also, Predictions were performed on the testing datasets which shows good accuracy when comparing actual versus predicted. Figure 5.13 shows a snippet of code results showing actual versus predicted energy in kwh and the residual difference.Moreover, figure 5.14 shows plotted results.

Actual KWH	Predicted KWH	Difference
8747.647	8445.533330	302.113670
8522.000	8979.400620	-457.400620
9020.942	9766.862287	-745.920287
5164.879	5690.463656	-525.584656
9553.608	8689.022130	864.585870
17029.688	16741.776304	287.911696
5610.048	5369.903889	240.144111
10541.384	10198.007608	343.376392
9135.233	8156.553670	978.679330
6989.000	7948.207870	-959.207870
8250.933	8792.871744	-541.938744
7867.981	8249.334755	-381.353755
7689.190	6736.788325	952.401675
19186.999	18677.378392	509.620608
4446.434	3803.050585	643.383415

Figure 5-13 kwh results

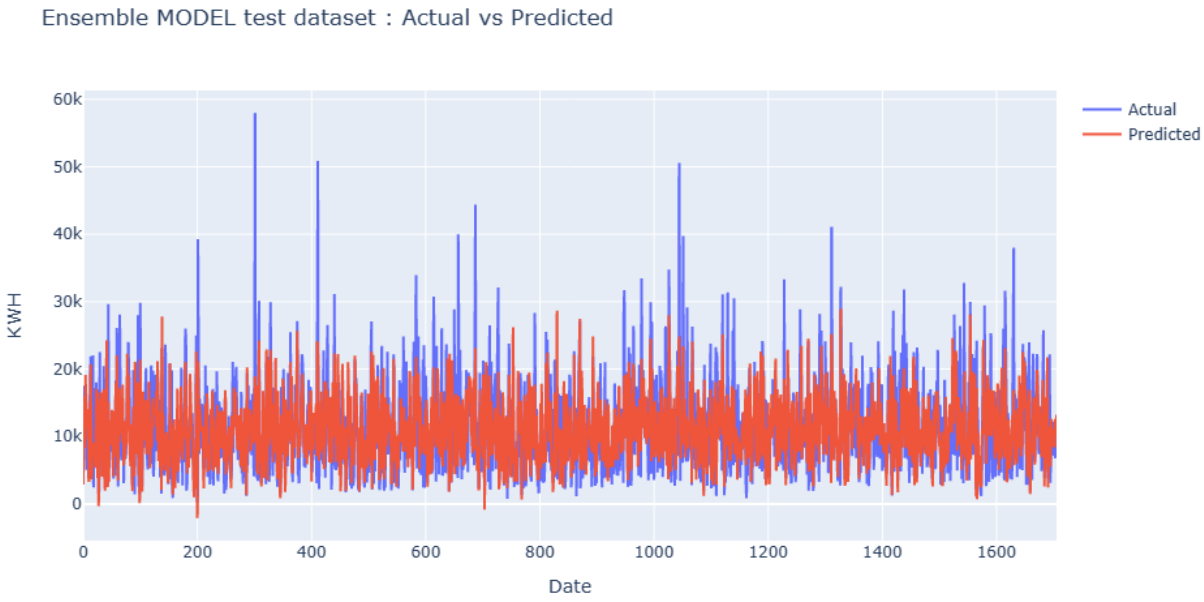


Figure 5-14 ensemble model actual vs predicted graph

5.4. conclusion

This model development findings indicate that employing ensemble machine learning model, in combining Ridge, Lasso, and Gradient Boosting, and random forest can substantially enhance the accuracy of energy consumption regression models. Future work may focus on exploring other modeling

techniques, addressing limitations, and further refining the model to improve prediction accuracy.

This chapter answers the research question 3

"How to evaluate the accuracy prediction improvement of the proposed occupancy"

The model is evaluated using a set of metrics and performed predictions which were compared to actual energy use.

AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE:
CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

MODEL VALIDATION
Chapter 6

CHAPTER 6

MODEL VALIDATION

6.1 Introduction

In this chapter, we discuss the model validation process of our ensemble model through a real case study. The case study involves the simulation and prediction of energy use in a residential building through design builder and energy plus software, and the results are compared against the actual energy performance of the building and the ensemble model.

The purpose of this case study is to evaluate the accuracy and effectiveness of our ensemble model in predicting energy use in a real-world scenario. To do so, we first collected data on various features that could impact energy use in the building, including occupancy behavior, building characteristics. The data is then used as an input to our ensemble model, which combines the predictions of several regression algorithms to achieve higher accuracy and lower errors.

To validate our model, we conducted a simulation of the building's energy use based on the collected data and compared it to the predictions of our ensemble model. We also compared our model's predictions to the actual energy performance of the building to evaluate its accuracy.

The results of our validation process showed that our ensemble model outperformed the simulation in predicting energy use in the building.

These results suggest that our ensemble model is an effective tool for predicting energy use in residential buildings, and can provide more accurate and reliable predictions than traditional simulation models and regression algorithms. This has significant implications for building owners, managers, and policymakers, as it can help identify opportunities for energy savings and

inform decision-making related to building energy use. Overall, our case study provides evidence for the effectiveness of ensemble models in energy prediction, and highlights the importance of model validation in ensuring accurate and reliable predictions.

This chapter concludes with answers for our research questions and validation of our hypothesis.

6.2 Casestudy approach

6.2.1 Case study selection

The selection of the case study was based on its suitability to our dataset, as well as the availability of relevant parameters that could be used to develop the design-builder, such as layouts and schedules. The chosen case study was adapted from the study conducted by (Yoshino et al., 2017). The figures below provide detailed information on the case study including a description of the case study and heating power schedule as shown in figure 6.1 and figure 6.2.

Category: R1
Data level: Complex level
Location: Sendai, Japan
Number of floors: 2
Heated area: 285 m^2
Construction year: 2008

Figure 6-1 case study description

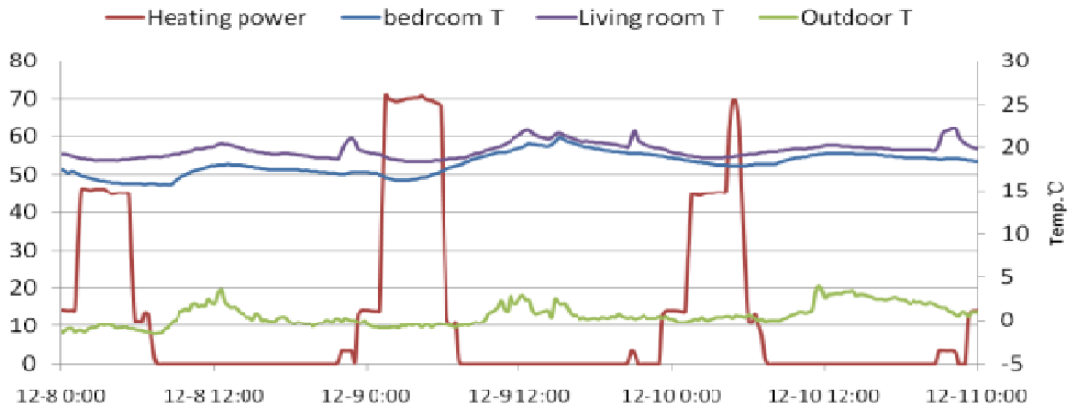


Figure 6-2 heating schedule

6.2.2 Simulation model parameters and process

In the process of simulating a case study, there were challenges related to privacy concerns regarding occupant tracking and identification. However, our case study includes occupant behavior and energy results was adequate to carry simulation and energy prediction.

The simulation process involved several steps. First, the building geometry was developed (Fig 6.3).

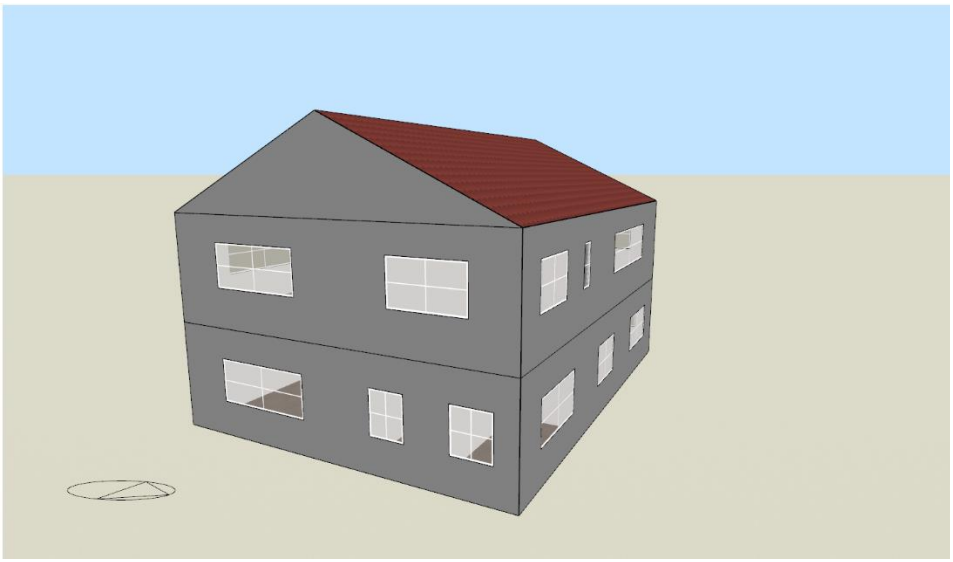


Figure 6-3 building geometry

The number of occupants, their schedule, and metabolic rate were provided. The cooling and heating set points, equipment power density and schedule, construction template, window layout and type, lighting power density and schedule, HVAC type, HVAC schedule were all adjusted to reflect the case study and suit the simulation requirements. Figure 6.4 provides a snippet from designbuilder parameter input.

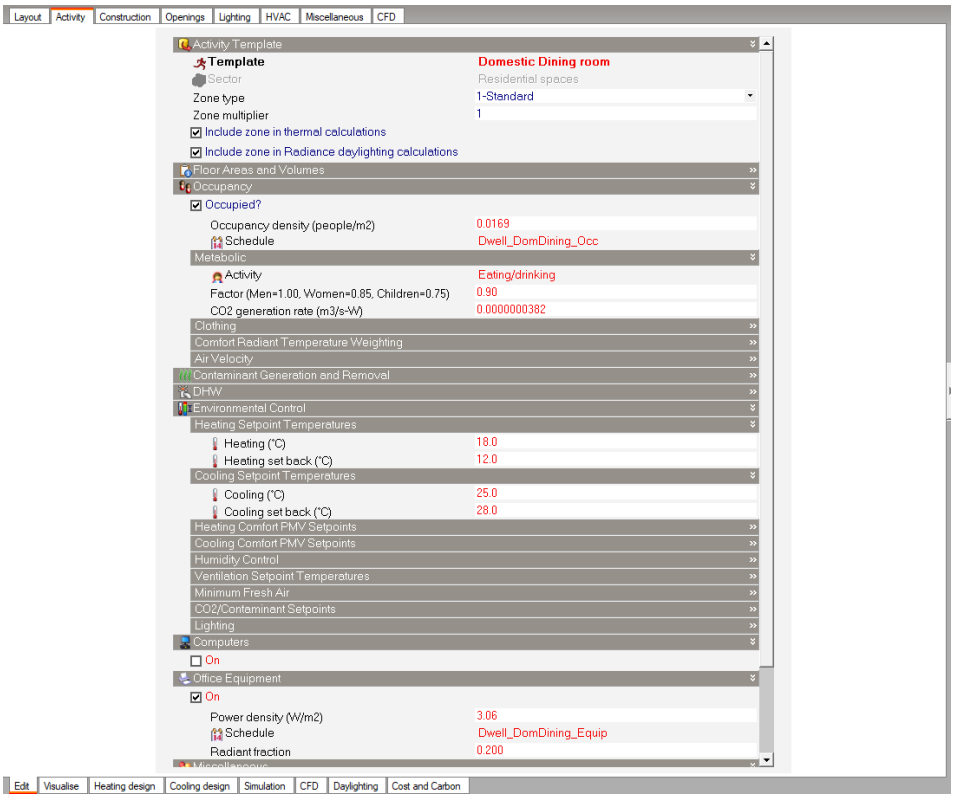


Figure 6-4 designbuilder inputs

EnergyPlus is a comprehensive building energy simulation program that relies on a detailed set of input data to model the energy dynamics of a building. The inputs for EnergyPlus can be quite extensive and are typically organized into an Input Data File (IDF). Below is a simplified table 6.1 that outlines some of the key categories and inputs required for a simulation in EnergyPlus and how these categories compares to the ensemble model input parameters presented in the next section.

This table provides an overview of the types of inputs that are typically required for an EnergyPlus simulation. Each category can have many specific parameters that need to be defined based on the building being modeled. The inputs are highly detailed to allow for an accurate simulation of the building's energy performance under various conditions.

In summary, the simulation process involved adjusting various building parameters to achieve energy prediction by running the simulation on energy plus. By using the appropriate tools and settings, the simulation was able to provide an energy prediction for the case study.

Table 6.1
Parameters of the model versus simulation

Design Builder Parameter	Value	Comparison to ML model parameter
Building geometry	Based on layout	Total square footage (used for publication) Total number of rooms in the housing unit, Number of sliding glass doors, windows..
Number of people	0.0169 people/m2; calculated based on numbers given	Number of household members
Occupant Schedule	Different for each zone, based on TM59	Number of weekdays someone is at home
Metabolic rate	110 W/person; default input	Number of household members age 18 or older Number of household members age 17 or younger
Cooling set point / setback point	25/28, default input	N/A
Heating set point / setback point	18/12; default input	N/A
Equipment power density	3.06 W/m2; default input	Number of refrigerators used Number of separate freezers used Number of stoves Number of separate cooktops Number of separate ovens Microwave oven used Number of desktop computers Number of laptop computers Number of tablet computers or e-readers Internet access at home Have dishwasher Have clothes washer in home Have clothes dryer in home Number of televisions used
Equipment schedule	Different for each zone, based on TM59	Frequency hot meals are cooked Frequency of dishwasher use Frequency of clothes washer use Frequency of clothes dryer use Most-used TV usage on weekdays Most-used TV usage on weekends Frequency of use of oven part of stove

Lighting power density	2.5116 W/m2; default input	Number of light bulbs installed inside the home
Lighting schedule	Different for each zone, based on TM59	Number of inside light bulbs turned on at least 4 hours a day
HVAC type	Heat pump	Space heating used Main space heating equipment type Any thermostats Programmable main thermostat Air conditioning equipment used Type of air conditioning equipment used Thermostat for central air conditioner Programmable thermostat for central air conditioner Number of whole house fans used Fuel used by main water heater
HVAC schedule	Different for each zone, based on TM59	Space heating usage Main heating equipment household behavior Central air conditioner household behavior

6.2.3 Ensemble model parameters and execution

In order to run the data derived from the case study in our model the following steps are followed. Using pandas library, the data is imported from a CSV file where the parameters of the input variables are saved (Table 6.2).

Table 6.2 Ensemble model input

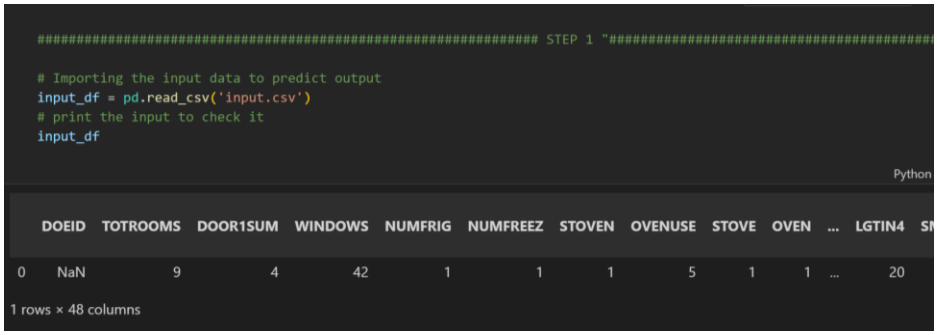
Parameter	Value	Category	Description	Source
TOTROOMS	9	Geometry	Total number of rooms	dataset
DOOR1SUM	4	Geometry	Number of doors	layout
WINDOWS	42	Geometry	Number of windows	layout
NUMFRIG	1	Appliances	Number of refrigerators	layout
NUMFREEZ	1	Appliances	Number of freezers	layout
STOVEN	1	Appliances	Number of stove-oven combinations	layout
OVENUSE	5	Behavioral	Frequency of oven use	inferred
STOVE	1	Appliances	Number of standalone stoves	layout
OVEN	1	Appliances	Number of standalone ovens	layout
MICRO	1	Appliances	Number of microwaves	layout
NUMMEAL	7	Behavioral	Frequency of hot meals cooked	inferred
DISHWASH	0	Appliances	Number of dishwashers	layout
DWASHUSE	0	Behavioral	Frequency of dishwasher use	inferred

CWASHER	1	Appliances	Number of washing machines	layout
WASHLOAD	4	Behavioral	Frequency of wash loads	inferred
DRYER	1	Appliances	Number of dryers	layout
DRYRUSE	4	Behavioral	Frequency of dryer use	inferred
TVCOLOR	3	Appliances	Number of color TVs	inferred
TVONWD1	2	Behavioral	TV usage on weekdays	inferred
TVONWE1	3	Behavioral	TV usage on weekends	inferred
DESKTOP	0	Appliances	Number of desktop computers	layout
NUMLAPTOP	3	Appliances	Number of laptop computers	inferred
NUMTABLET	5	Appliances	Number of tablet computers	inferred
INTERNET	1	System	Presence of internet access	inferred
HEATHOME	1	System	Space heating usage	dataset
EQUIPM	1	System	Age of main space heating equipment	inferred
THERMAIN	1	System	Presence of any thermostats	dataset
PROTHERM	1	System	Presence of programmable main thermostat	dataset
EQUIPMUSE	2	Behavioral	Main heating equipment household behavior	inferred
AIRCOND	1	System	Air conditioning equipment usage	inferred
COOLTYPE	2	System	Type of air conditioning equipment used	Given
THERMAINAC	0	System	Thermostat for central air conditioner	dataset
PROTHERMAC	1	System	Programmable thermostat for central air conditioner	dataset
USECENAC	1	System	Central air conditioner household behavior	inferred
NUMWHOLEFAN	0	Appliances	Number of whole house fans	layout
FUELH2O	5	System	Fuel used by main water heater	dataset
LGTINNUM	20	Lighting	Number of light bulbs installed inside the home	layout
LGTIN4	20	Behavioral	Number of inside light bulbs turned on at least 4 hours a day	inferred
SMARTTHERM	1	System	Presence of smart thermostat	inferred
SMARTMETER	1	System	Presence of smart meter	dataset
NHSLDMEM	5	Demographics	Number of household members	dataset
NUMADULT	2	Demographics	Number of adults in the household	inferred
NUMCHILD	3	Demographics	Number of children in the household	inferred

ATHOME	5	Behavioral	Number of weekdays someone is at home	inferred
Total Area sqf	3067	Geometry	Total area in square feet	dataset

The data is then combined with the training and testing data from the Ashrae dataset. Categorical variables are encoded using dummy encoding technique. The necessary transformations are applied to the input to fit the scaler and ensemble model. Logarithmic transformation is applied to the same columns as the Ashrae dataset. The 'id' column is dropped, and the same scaler is used to transform the input. Finally, the trained regression model is used to predict the energy consumption output for the given input data using the 'predict' method.

Code snippet is presented below showing the data input in figure 6.5, data combination with the existing dataset in figure 6.6, and the output prediction value in figure 6.7.



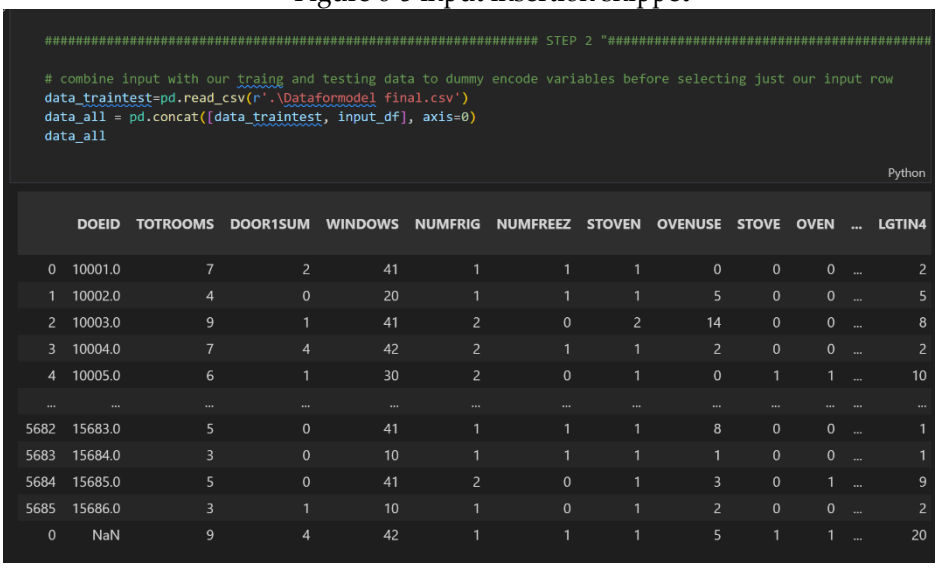
```
############################################################ STEP 1 "#########################################

# Importing the input data to predict output
input_df = pd.read_csv('input.csv')
# print the input to check it
input_df
```

Python

	DOEID	TOTROOMS	DOOR1SUM	WINDOWS	NUMFRIG	NUMFREEZ	STOVEN	OVENUSE	STOVE	OVEN	...	LGTIN4	SI
0	NaN	9	4	42	1	1	1	5	1	1	...	20	

1 rows × 48 columns

Figure 6-5 input insertion snippet



```
############################################################ STEP 2 "#########################################

# combine input with our traing and testing data to dummy encode variables before selecting just our input row
data_traintest=pd.read_csv(r'.\Dataformodel final.csv')
data_all = pd.concat([data_traintest, input_df], axis=0)
data_all
```

Python

	DOEID	TOTROOMS	DOOR1SUM	WINDOWS	NUMFRIG	NUMFREEZ	STOVEN	OVENUSE	STOVE	OVEN	...	LGTIN4
0	10001.0	7	2	41	1	1	1	0	0	0	...	2
1	10002.0	4	0	20	1	1	1	5	0	0	...	5
2	10003.0	9	1	41	2	0	2	14	0	0	...	8
3	10004.0	7	4	42	2	1	1	2	0	0	...	2
4	10005.0	6	1	30	2	0	1	0	1	1	...	10
...	...	...	...	...	...	...	...	...	...	...	...	...
5682	15683.0	5	0	41	1	1	1	8	0	0	...	1
5683	15684.0	3	0	10	1	1	1	1	0	0	...	1
5684	15685.0	5	0	41	2	0	1	3	0	1	...	9
5685	15686.0	3	1	10	1	0	1	2	0	0	...	2
0	NaN	9	4	42	1	1	1	5	1	1	...	20

Figure 6-6 combining datasets snippet

```
# Use the model and predict method and get the output
Best_Model.predict(X_1)
```

```
array([2100.52944141])
```

Figure 6-7 prediction output snippet

6.3 Comparative analysis and results

6.3.1 Input comparison and constraints

In terms of building energy performance, the approaches for simulation and predictive modeling can vary significantly, particularly in terms of input requirements. EnergyPlus simulations typically rely on a standardized set of inputs that define the physical and operational characteristics of a building. These inputs can often be set to default values when specific data is not available, ensuring that the simulation can still proceed with reasonable assumptions about typical building behavior.

For the ensemble machine learning model developed in this study, the input parameters are distinct, especially concerning occupant behavior. The model leverages a set of occupant-related inputs, as specified in table 6.2, which includes variables such as oven and stove usage, meal frequency, appliance use, and occupancy patterns. These inputs are crucial as they directly influence the energy consumption patterns within the residential setting.

The data for these occupant variables is sourced from the case study or deduced based on the information provided therein. When the case study lacks specific details, we resort to making educated assumptions. These assumptions are informed by historical data and similar case studies, which have been analyzed through machine learning techniques. Our model benefits from being trained on extensive datasets, which permeate it with the capability to infer or predict inputs with a degree of confidence.

This approach allows for a more nuanced understanding of occupant behavior and its impact on energy performance. While the EnergyPlus simulation provides a baseline by using standard and default inputs, our ensemble model

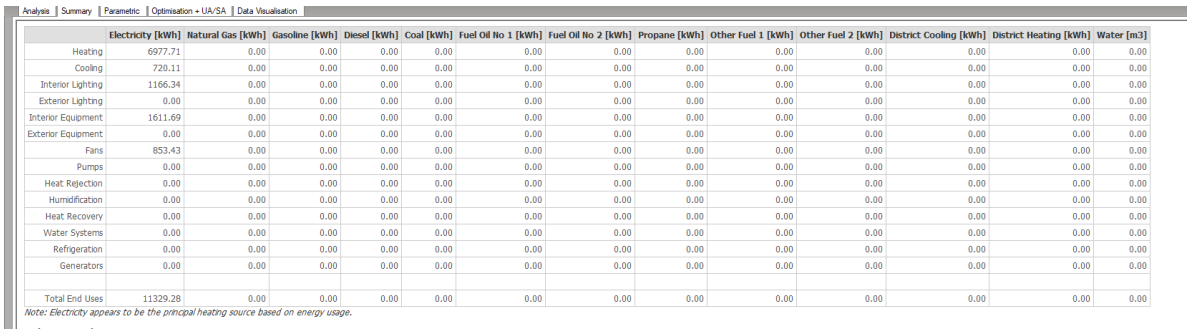
has a layer of occupant behavior complexity, offering predictions that are potentially more aligned with the actual energy usage patterns.

By integrating these diverse data sources and leveraging the predictive power of machine learning, the ensemble model aims to reduce the gap between predicted and actual energy consumption. This is particularly important in the context of residential buildings, where occupant behavior is a significant and often variable component of energy use.

6.3.2 Results comparison

The energy consumption results were obtained by executing both the simulation model and ensemble model. The actual energy consumption was then compared against these results. Upon analysis, it was found that the actual energy consumption was approximately 11329 kwh/m2/a when using the simulation model (figure 6.8).

On the other hand, the ensemble model predicted the energy consumption to be around 2100.5 kwh/m2/a (figure 6.7). However, the actual energy consumption turned out to be approximately 2400 kwh/m2/a (figure 6.9).



Analysis | Summary | Parametric | Optimisation + UA/SA | Data Visualisation

	Electricity [kWh]	Natural Gas [kWh]	Gasoline [kWh]	Diesel [kWh]	Coal [kWh]	Fuel Oil No 1 [kWh]	Fuel Oil No 2 [kWh]	Propane [kWh]	Other Fuel 1 [kWh]	Other Fuel 2 [kWh]	District Cooling [kWh]	District Heating [kWh]	Water [m3]
Heating	6977.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cooling	720.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Interior Lighting	1166.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Exterior Lighting	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Interior Equipment	1611.69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Exterior Equipment	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fans	853.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pumps	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Heat Rejection	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Humidification	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Heat Recovery	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Water Systems	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Refrigeration	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Generators	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total End Uses	11329.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: Electricity appears to be the principal heating source based on energy usage.

Figure 6-8 simulation results

Based on the results obtained from this case study, it can be concluded that the ensemble model developed in this study performed well in predicting the energy consumption of residential buildings. This can be seen from the comparison of the actual energy consumption with the predicted energy consumption by the ensemble model. This suggests that the ensemble model can be developed and used as a reliable tool for predicting energy consumption in residential buildings, which can be useful for energy efficiency improvement and cost reduction. This in return leads to closing the gap

between actual and predicted energy performance while accounting for occupant behaviors parameters.

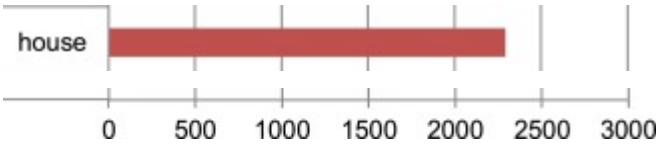


Figure 6-9 actual energy consumption

It is plausible that the simulation's assumptions, particularly regarding occupant behavior and equipment schedules, were not reflective of the actual conditions. This misalignment could be a contributing factor to the simulation's poor performance. The ensemble model, on the other hand, benefits from a more dynamic input set that accounts for occupant behavior, which is a critical determinant of energy consumption in residential buildings.

6.4 Limitations and future work

A limitation of our study is that assumptions had to be made for certain parameters related to occupant behaviors in the ensemble model since the values were not available in the case study literature. However, based on the results of this particular case study, our model performed well. It is important to note that since the dataset is limited and access to a larger number of case studies is not possible, we need to build more confidence in our model's accuracy as we gather more data.

Another limitation of our study is that the ensemble model is based on machine learning algorithms that require a large amount of data to train and optimize. Since we only used one case study to train our model, there is a possibility that the model may not generalize well to other buildings with different characteristics.

In addition, our study only focused on energy consumption prediction and did not take into account other important factors such as indoor environmental quality and occupant comfort. Future studies can explore the integration of these factors into the model to provide a more comprehensive analysis of building performance.

Another area for future work is to incorporate real-time data from building automation systems and IoT devices to improve the accuracy of the model. This would allow for more dynamic and responsive predictions of energy consumption and occupant behavior.

Finally, it is important to continue to validate and refine the model with more case studies from different regions and building types. This would help to improve the model's accuracy and robustness, as well as increase its applicability to a wider range of buildings and contexts.

6.5 Conclusion

In conclusion, the validation of the ensemble model against the case study and the simulation from EnergyPlus reveals its superior predictive accuracy. However, it is imperative to acknowledge the limitations that may have influenced the validation outcomes. The assumptions made for certain parameters, especially those pertaining to occupant behavior, could have introduced a degree of uncertainty in the simulation model predictions. Future work should aim to minimize these assumptions by incorporating real-time data and expanding the dataset to include a more diverse range of case studies. This would not only enhance the model's accuracy but also its generalizability across different building types and occupant profiles.

AN ENSEMBLE MODEL FOR PREDICTIVE ENERGY PERFORMANCE:
CLOSING THE GAP BETWEEN ACTUAL AND PREDICTED ENERGY USE IN RESIDENTIAL BUILDINGS

CONCLUSION
Chapter 7

CHAPTER 7

CONCLUSION

7.1 Summary

This research embarked on developing an ensemble machine learning model to predict residential building energy consumption accurately. The journey began with defining specific objectives aimed at addressing the existing gap between predicted and actual energy use, with a keen focus on integrating occupant behavior into the predictive models.

To achieve these objectives, the following measures were taken:

- **Objective 1:** Identification of Influential Parameters
  Through a systematic review and evaluation of machine learning algorithms, we identified critical occupant-related parameters influencing residential energy performance. An extensive dataset was used to analyze and understand the correlation between these parameters and energy consumption.
- **Objective 2:** Development of the Ensemble Model
  An ensemble machine learning model incorporating Lasso regression, Ridge regression, Random Forest, and Gradient Boosting was developed. This model was tailored to include a comprehensive set of occupant behavior parameters to enhance prediction accuracy.
- **Objective 3:** Improvement of Prediction Accuracy
  The model's accuracy was rigorously evaluated by comparing it against solo prediction models. The ensemble approach significantly improved prediction accuracy, validating the hypothesis that a machine learning model informed by occupant behavior can outperform traditional methods.
- **Objective 4:** Validation of Model Reliability and Applicability

The final validation utilized a case study to demonstrate the real-world applicability of the ensemble model. The model's predictions were compared to actual energy consumption data, confirming its reliability and practical value.

The findings revealed that the ensemble model effectively reduces the gap between predicted and actual energy use, thereby validating the research hypothesis. Notably, the model demonstrated superior performance over traditional simulation method using designbuilder and energyplus, particularly in capturing the nuances of occupant behavior.

In conclusion, the ensemble machine learning model represents a significant advancement in predictive accuracy for residential building energy consumption. It stands as a testament to the potential of machine learning in transforming the field of energy modeling, offering a pathway to more sustainable building management practices. The successful validation of the model reaffirms the value of incorporating a broad spectrum of occupant behavior into predictive models, which can significantly impact energy conservation strategies and policies globally.

this research aimed to develop an occupancy behavior-based ensemble machine learning model to reduce the gap between actual and predicted energy consumption in residential buildings by integrating all factors contributing to occupants' behaviors into building energy predictions. The inclusion of occupant behavior impact on energy consumption was the primary focus of this research, as it sought to minimize the energy performance gap and provide more reliable predictions.

In line with the research objectives, this study has successfully developed an ensemble machine learning predictive model for residential building energy consumption that incorporates occupant behavior-based inputs. The model was evaluated and validated, demonstrating its potential for improving the accuracy of energy consumption predictions and fostering better energy management strategies. The findings of this research can provide valuable guidance for energy modelers, building designers, and policymakers seeking to promote energy efficiency in residential buildings while considering the complex and often unpredictable nature of occupant behavior.

7.2 Conclusions

The study's conclusions are manifold:

The ensemble model showcased an advantage in predicting residential energy energy performance by incorporating detailed occupant behavior data.

The predictive accuracy of the ensemble model was systematically validated against actual energy use, underscoring its practical applicability in real-world settings.

This research contributes to the body of knowledge by providing a robust methodological framework for developing energy prediction models that integrate diverse and dynamic parameters affecting residential energy use.

This study's approach addresses the critical challenge of reducing the energy performance gap, offering a tool that can enhance energy efficiency measures and sustainability in the residential sector. It also lays the groundwork for incorporating machine learning into building energy management, steering towards more intelligent and adaptive systems.

The generalizability of these findings is promising, given the model's capacity to accommodate various building types and occupant profiles. It paves the way for deploying similar models in diverse contexts, contributing to global efforts in energy conservation and climate change mitigation.

7.3 Limitations

Despite the model's success, the research encountered limitations:

- The ensemble model's predictive accuracy is contingent on the richness of the dataset. In cases where detailed occupant behavior data were not available, assumptions were made, which may limit the model's precision.
- The study's scope was restricted to a single case study for validation, which may not fully represent the diversity of residential settings.
- The current model focuses solely on energy consumption, omitting other relevant aspects such as thermal comfort and indoor environmental quality.

7.4 Future Research

Future research directions include:

- Expanding the dataset to include a wider array of residential settings, occupant behaviors, and climate zones to enhance the model's robustness and generalizability.
- Integrating real-time data from IoT devices and building management systems to capture dynamic changes in occupant behavior and environmental conditions.
- Including other dimensions of building performance in the model, such as indoor air quality and occupant comfort, to develop a more holistic tool for building energy management.
- Applying the model in longitudinal studies to verify its performance over longer periods and under various seasonal conditions.

REFERENCES

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and buildings*, *147*, 77-89.

Alaaeddine, R., & Wu, S. (2017). Application of supervised learning methods to better predict building energy performance.

Andersen, R. V., Olesen, B. W., & Toftum, J. (2011). Modelling occupants' heating set-point preferences. Building Simulation Conference,

ASHRAE. (2018). *ASHRAE global occupant behavior database*. American Society of Heating, Refrigerating and Air-Conditioning Engineers. https://www.ashrae.org/professional-development/occupant-behavior-database

ASHRAE. (2021). *ASHRAE Global Thermal Comfort Database II*. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). https://www.ashrae.org/technical-resources/global-thermal-comfort-database-ii

Azar, E., Syndicus, M., Markovic, R., Alsereidi, A., Wagner, A., Frisch, J., & van Treeck, C. (2022). Crossing borders and methods: Comparing individual and social influences on energy saving in the United Arab Emirates and Germany. *Energy Research & Social Science*, *90*, 102561.

Barthelmes, V. M., Becchio, C., Fabi, V., & Corgnati, S. P. (2017). Occupant behaviour lifestyles and effects on building energy use: Investigation on high and low performing building features. *Energy Procedia*, *140*, 93-101.

Barthelmes, V. M., Heo, Y., Fabi, V., & Corgnati, S. P. (2017). Exploration of the Bayesian Network framework for modelling window control behaviour. *Building and Environment*, *126*, 318-330.

Basu, K., Hawarah, L., Arghira, N., Joumaa, H., & Ploix, S. (2013). A prediction system for home appliance usage. *Energy and Buildings*, *67*, 668-679.

Becchio, C., Bello, C., Corgnati, S., & Ingaramo, L. (2016). Influence of occupant behaviour lifestyle on an Italian social housing. *Energy Procedia*, *101*, 1034-1041.

Borgstein, E., Lamberts, R., & Hensen, J. (2016). Evaluating energy performance in non-domestic buildings: A review. *Energy and Buildings*, *128*, 734-755.

Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2018). Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, *11*(7), 1636.

Bruce-Konuah, A. (2014). *Occupant window opening behaviour: the relative importance of temperature and carbon dioxide in university office buildings* University of Sheffield].

Buratti, C., Belloni, E., & Palladino, D. (2014). Evolutive Housing System: Refurbishment with new technologies and unsteady simulations of energy performance. *Energy and Buildings*, *74*, 173-181.

Buso, T., Fabi, V., Andersen, R. K., & Corgnati, S. P. (2015). Occupant behaviour and robustness of building design. *Building and Environment*, *94*, 694-703.

Calì, D., Andersen, R. K., Müller, D., & Olesen, B. W. (2016). Analysis of occupants' behavior related to the use of windows in German households. *Building and Environment*, *103*, 54-69.
Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, *140*, 81-97.
Cao, X., Dai, X., & Liu, J. (2016). Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. *Energy and buildings*, *128*, 198-213.
Castaldo, V. L., & Pisello, A. L. (2018). Uses of dynamic simulation to predict thermal-energy performance of buildings and districts: a review. *Wiley Interdisciplinary Reviews: Energy and Environment*, *7*(1), e269.
Cetin, K., Tabares-Velasco, P., & Novoselac, A. (2014). Appliance daily energy use in new residential buildings: Use profiles and variation in time-of-use. *Energy and Buildings*, *84*, 716-726.
Chitalia, G., Pipattanasomporn, M., Garg, V., & Rahman, S. (2020). Robust short-term electrical load forecasting framework for commercial buildings using deep recurrent neural networks. *Applied Energy*, *278*, 115410.
Conner, M., & Armitage, C. J. (1998). Extending the theory of planned behavior: A review and avenues for further research. *Journal of applied social psychology*, *28*(15), 1429-1464.
Crawley, D. B., Lawrie, L. K., Pedersen, C. O., & Winkelmann, F. C. (2000). Energy plus: energy simulation program. *ASHRAE journal*, *42*(4), 49-56.
D'Oca, S., Fabi, V., Corgnati, S. P., & Andersen, R. K. (2014). Effect of thermostat and window opening occupant behavior models on energy use in homes. Building Simulation,
da Silva, P. C., Leal, V., & Andersen, M. (2013). Occupants interaction with electric lighting and shading systems in real single-occupied offices: Results from a monitoring campaign. *Building and Environment*, *64*, 152-168.
Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge University Press.
Daum, D., & Morel, N. (2010). Assessing the total energy impact of manual and optimized blind control in combination with different lighting schedules in a building simulation environment. *Journal of Building Performance Simulation*, *3*(1), 1-16.
de Dear, R. J., Akimoto, T., Arens, E. A., Brager, G., Candido, C., Cheong, K., Li, B., Nishihara, N., Sekhar, S., & Tanabe, S. (2013). Progress in thermal comfort research over the last twenty years. *Indoor air*, *23*(6), 442-461.
De Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in Construction*, *41*, 40-49.
Delzendeh, E., Wu, S., Lee, A., & Zhou, Y. (2017). The impact of occupants' behaviours on building energy analysis: A research review. *Renewable and Sustainable Energy Reviews*, *80*, 1061-1071. https://doi.org/https://doi.org/10.1016/j.rser.2017.05.264
Demanuele, C., Tweddell, T., & Davies, M. (2010). Bridging the gap between predicted and actual energy performance in schools. World renewable energy congress XI,
Deng, H., Fannon, D., & Eckelman, M. J. (2018). Predictive modeling for US commercial building energy use: A comparison of existing statistical and

machine learning algorithms using CBECS microdata. *Energy and Buildings*, *163*, 34-43.
Deng, Z., & Chen, Q. (2018). Artificial neural network models using thermal sensations and occupants' behavior for predicting thermal comfort. *Energy and Buildings*, *174*, 587-602.
Ding, Y., Fan, L., & Liu, X. (2021). Analysis of feature matrix in machine learning algorithms to predict energy consumption of public buildings. *Energy and Buildings*, *249*, 111208.
Divina, F., Torres Maldonado, J. F., García-Torres, M., Martínez-Álvarez, F., & Troncoso, A. (2020). Hybridizing deep learning and neuroevolution: application to the Spanish short-term electric energy consumption forecasting. *Applied Sciences*, *10*(16), 5487.
Dong, J., Schwartz, Y., Mavrogianni, A., Korolija, I., & Mumovic, D. (2023). A review of approaches and applications in building stock energy and indoor environment modelling. *Building Services Engineering Research and Technology*, 01436244231163084.
Dong, Z., Liu, J., Liu, B., Li, K., & Li, X. (2021). Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification. *Energy and Buildings*, *241*, 110929.
Dutton, S., & Shao, L. (2010). Window opening behaviour in a naturally ventilated school. *Proceedings of SimBuild*, *4*(1), 260-268.
EIA. (2021). *Commercial Buildings Energy Consumption Survey (CBECS)*. U.S. Energy Information Administration (EIA). Retrieved 2021 from https://www.eia.gov/consumption/commercial/
Fabi, V., Andersen, R. K., & Corgnati, S. (2016). Accounting for the uncertainty related to building occupants with regards to visual comfort: A literature survey on drivers and models. *Buildings*, *6*(1), 5.
Fabi, V., Andersen, R. V., Corgnati, S., & Olesen, B. W. (2012). Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models. *Building and Environment*, *58*, 188-198.
Fabi, V., Andersen, R. V., Corgnati, S. P., & Olesen, B. W. (2013). A methodology for modelling energy-related human behaviour: Application to window opening behaviour in residential buildings. Building Simulation,
Fabi, V., Andersen, R. V., Corgnati, S. P., Olesen, B. W., & Filippi, M. (2011). Description of occupant behaviour in building energy simulation: state-of-art and concepts for improvements. Proceedings of building simulation,
Fabi, V., Camisassi, V., Causone, F., Corgnati, S., & Andersen, R. (2014). Light switch behaviour: occupant behaviour stochastic models in office buildings. 8th windsor conference,
Farghali, M., Osman, A. I., Mohamed, I. M., Chen, Z., Chen, L., Ihara, I., Yap, P.-S., & Rooney, D. W. (2023). Strategies to save energy in the context of the energy crisis: a review. *Environmental Chemistry Letters*, 1-37.
Fayaz, M., & Kim, D. (2018). A prediction methodology of energy consumption based on deep extreme learning machine and comparative analysis in residential buildings. *Electronics*, *7*(10), 222.
Fekri, M. N., Patel, H., Grolinger, K., & Sharma, V. (2021). Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network. *Applied Energy*, *282*, 116177.

Foucquier, A., Robert, S., Suard, F., Stéphan, L., & Jay, A. (2013). State of the art in building modelling and energy performances prediction: A review. *Renewable and Sustainable Energy Reviews*, *23*, 272-288.

Ghahramani, A., Karvigh, S. A., & Becerik-Gerber, B. (2017). HVAC system energy optimization using an adaptive hybrid metaheuristic. *Energy and Buildings*.

Gunay, H. B., O'Brien, W., Beausoleil-Morrison, I., & Bursill, J. (2018). Development and implementation of a thermostat learning algorithm. *Science and Technology for the Built Environment*, *24*(1), 43-56.

Gunay, H. B., O'Brien, W., Beausoleil-Morrison, I., & Gilani, S. (2017). Development and implementation of an adaptive lighting and blinds control algorithm. *Building and Environment*, *113*, 185-199.

Gunay, H. B., O'Brien, W., Beausoleil-Morrison, I., & Huchuk, B. (2014). On adaptive occupant-learning window blind and lighting controls. *Building Research & Information*, *42*(6), 739-756.

Guo, J., Yun, S., Meng, Y., He, N., Ye, D., Zhao, Z., Jia, L., & Yang, L. (2023). Prediction of heating and cooling loads based on light gradient boosting machine algorithms. *Building and Environment*, 110252.

Haldi, F., Calì, D., Andersen, R. K., Wesseling, M., & Müller, D. (2017). Modelling diversity in building occupant behaviour: a novel statistical approach. *Journal of Building Performance Simulation*, *10*(5-6), 527-544.

Haldi, F., & Robinson, D. (2010). Adaptive actions on shading devices in response to local visual stimuli. *Journal of Building Performance Simulation*, *3*(2), 135-153.

Haldi, F., & Robinson, D. (2011). The impact of occupants' behaviour on building energy demand. *Journal of Building Performance Simulation*, *4*(4), 323-338.

Harputlugil, T., & de Wilde, P. (2021). The interaction between humans and buildings for energy efficiency: A critical review. *Energy Research & Social Science*, *71*, 101828.

Hasselbring, W., & Giesecke, S. (2006). Research Methods in Software Engineering. GITO-Verl.

He, Y., Qin, Y., Wang, S., Wang, X., & Wang, C. (2019). Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network. *Applied energy*, *233*, 565-575.

Heating, A. S. o., Refrigerating, & Engineers, A.-C. (2000). *Heating, Ventilating, and Air-Conditioning: Systems and Equipment: 2000 Ashrae Handbook: Inch-Pound*. Amer Society of Heating.

Higginson, S., McKenna, E., Hargreaves, T., Chilvers, J., & Thomson, M. (2015). Diagramming social practice theory: An interdisciplinary experiment exploring practices as networks. *Indoor and Built Environment*, *24*(7), 950-969.

Hong, T., D'Oca, S., Turner, W. J., & Taylor-Lange, S. C. (2015). An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to the DNAs framework. *Building and Environment*, *92*, 764-777.

Hong, T., & Lin, H.-W. (2013). *Occupant behavior: impact on energy use of private offices*.

Hong, T., Taylor-Lange, S. C., D'Oca, S., Yan, D., & Corgnati, S. P. (2016). Advances in research and applications of energy-related occupant behavior in buildings. *Energy and Buildings*, *116*, 694-702.

Hong, T., Yan, D., D'Oca, S., & Chen, C.-f. (2017). Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment*, *114*, 518-530.
Huang, Y., Lu, T., Ding, X., & Gu, N. (2014). Campus Building Energy Usage Analysis and Prediction: A SVR Approach Based on Multi-scale RBF Kernels. International Conference on Human Centered Computing,
Huchuk, B., Gunay, H. B., O'Brien, W., & Cruickshank, C. A. (2016). Model-based predictive control of office window shades. *Building Research & Information*, *44*(4), 445-455.
Huebner, G. M., Hamilton, I., Chalabi, Z., Shipworth, D., & Oreszczyn, T. (2015). Explaining domestic energy consumption–the comparative contribution of building factors, socio-demographics, behaviours and attitudes. *Applied energy*, *159*, 589-600.
IEA. (2017). World Energy Outlook 2017. https://www.iea.org/reports/world-energy-outlook-2017
IEA. (2021). *International Energy Outlook 2021*. https://www.eia.gov/outlooks/ieo/".
Indraganti, M., Ooka, R., & Rijal, H. B. (2015). Thermal comfort in offices in India: Behavioral adaptation and the effect of age and gender. *Energy and Buildings*, *103*, 284-295.
Jain, M., Singh, A., & Chandan, V. (2016). Non-intrusive estimation and prediction of residential ac energy consumption. Pervasive Computing and Communications (PerCom), 2016 IEEE International Conference on,
Jami, S., Forouzandeh, N., Zomorodian, Z. S., Tahsildoost, M., & Khoshbakht, M. (2021). The effect of occupant behaviors on energy retrofit: A case study of student dormitories in Tehran. *Journal of Cleaner Production*, *278*, 123556.
Jamil, F., Iqbal, N., Ahmad, S., & Kim, D. (2021). Peer-to-peer energy trading mechanism based on blockchain and machine learning for sustainable electrical power supply in smart grid. *IEEE Access*, *9*, 39193-39217.
Janda, K. B. (2011). Buildings don't use energy: people do. *Architectural science review*, *54*(1), 15-22.
Jang, H., & Kang, J. (2016). A stochastic model of integrating occupant behaviour into energy simulation with respect to actual energy consumption in high-rise apartment buildings. *Energy and Buildings*, *121*, 205-216.
Jia, M., Srinivasan, R. S., & Raheem, A. A. (2017). From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency. *Renewable and Sustainable Energy Reviews*, *68*, 525-540.
Jiang, J., Wang, C., Roth, T., Nguyen, C., Kamongi, P., Lee, H., & Liu, Y. (2021). Residential house occupancy detection: Trust-based scheme using economic and privacy-aware sensors. *IEEE Internet of Things Journal*, *9*(3), 1938-1950.
Jin, X.-B., Zheng, W.-Z., Kong, J.-L., Wang, X.-Y., Bai, Y.-T., Su, T.-L., & Lin, S. (2021). Deep-learning forecasting method for electric power load via attention-based encoder-decoder with bayesian optimization. *Energies*, *14*(6), 1596.
Jin, Y., Yan, D., Chong, A., Dong, B., & An, J. (2021). Building occupancy forecasting: A systematical and critical review. *Energy and Buildings*, *251*, 111345.

Jones, R. V., Fuertes, A., Gregori, E., & Giretti, A. (2017). Stochastic behavioural models of occupants' main bedroom window operation for UK residential buildings. *Building and Environment*, *118*, 144-158.
Kaminski, J. (2011). Diffusion of innovation theory. *Canadian Journal of Nursing Informatics*, *6*(2), 1-6.
Kavousian, A., Rajagopal, R., & Fischer, M. (2015). Ranking appliance energy efficiency in households: Utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings. *Energy and Buildings*, *99*, 220-230.
Kaygusuz, K. (2012). Energy for sustainable development: A case of developing countries. *Renewable and Sustainable Energy Reviews*, *16*(2), 1116-1126.
Kim, J., de Dear, R., Parkinson, T., & Candido, C. (2017). Understanding patterns of adaptive comfort behaviour in the Sydney mixed-mode residential context. *Energy and Buildings*, *141*, 274-283.
Kim, J., Zhou, Y., Schiavon, S., Raftery, P., & Brager, G. (2018). Personal comfort models: predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning. *Building and Environment*, *129*, 96-106.
Kim, T.-Y., & Cho, S.-B. (2019a). Particle swarm optimization-based CNN-LSTM networks for forecasting energy consumption. 2019 IEEE congress on evolutionary computation (CEC),
Kim, T.-Y., & Cho, S.-B. (2019b). Predicting residential energy consumption using CNN-LSTM neural networks. *Energy*, *182*, 72-81.
Kiprijanovska, I., Stankoski, S., Ilievski, I., Jovanovski, S., Gams, M., & Gjoreski, H. (2020). Houseec: Day-ahead household electrical energy consumption forecasting using deep learning. *Energies*, *13*(10), 2672.
Kitchenham, & Charters. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *2*.
Kodratoff, Y. (2014). *Introduction to machine learning*. Elsevier.
Koehler, C., Ziebart, B. D., Mankoff, J., & Dey, A. K. (2013). TherML: occupancy prediction for thermostat control. Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing,
Koulamas, C., Kalogeras, A., Pacheco-Torres, R., Casillas, J., & Ferrarini, L. (2017). Suitability analysis of modeling and assessment approaches in energy efficiency in buildings. *Energy and Buildings*.
Krstić, H., & Teni, M. (2017). Review of Methods for Buildings Energy Performance Modelling. IOP Conference Series: Materials Science and Engineering,
Kruusimägi, M., Sharples, S., & Robinson, D. (2018). A novel spatiotemporal home heating controller design: System emulation and field testing. *Building and Environment*, *135*, 10-30.
Le, T., Vo, M. T., Vo, B., Hwang, E., Rho, S., & Baik, S. W. (2019). Improving electric energy consumption prediction using CNN and Bi-LSTM. *Applied Sciences*, *9*(20), 4237.
Lei, L., Chen, W., Wu, B., Chen, C., & Liu, W. (2021). A building energy consumption prediction model based on rough set theory and deep learning algorithms. *Energy and Buildings*, *240*, 110886.
Li, N., Li, J., Fan, R., & Jia, H. (2015). Probability of occupant operation of windows during transition seasons in office buildings. *Renewable Energy*, *73*, 84-91.

Li, Z., Han, Y., & Xu, P. (2014). Methods for benchmarking building energy consumption against its past or intended performance: An overview. *Applied Energy*, *124*, 325-334.
Li, Z., & Huang, G. (2013). Re-evaluation of building cooling load prediction models for use in humid subtropical area. *Energy and Buildings*, *62*, 442-449.
Lim, J. H., & Yun, G. Y. (2017). Cooling Energy Implications of Occupant Factor in Buildings under Climate Change. *Sustainability*, *9*(11), 2039.
Mahdavi, A., Tahmasebi, F., & Kayalar, M. (2016). Prediction of plug loads in office buildings: Simplified and probabilistic methods. *Energy and Buildings*, *129*, 322-329.
Markovic, R., Grintal, E., Wölki, D., Frisch, J., & van Treeck, C. (2018). Window Opening Model using Deep Learning Methods. *arXiv preprint arXiv:1807.03610*.
Markovic, R., Wolf, S., Cao, J., Spinnräker, E., Wölki, D., Frisch, J., & van Treeck, C. (2017). Comparison of Different Classification Algorithms for the Detection of User's Interaction with Windows in Office Buildings. *Energy Procedia*, *122*, 337-342.
Mehta, C., Fung, A. S., Engineer-in-Training APEGA, M., & GA, L. (2013). A Case Study in Actual Building Performance and Energy Modeling with Real Weather Data. In: Ryerson University.
Menezes, A. C., Cripps, A., Bouchlaghem, D., & Buswell, R. (2012). Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap. *Applied Energy*, *97*, 355-364.
Menezes, A. C. K. d., Tetlow, R., Beaman, C. P., Bouchlaghem, D., Cripps, A., & Buswell, R. A. (2012). Assessing the impact of occupant behaviour on electricity consumption for lighting and small power in office buildings.
Mohamed, A. R., & Lee, K. T. (2006). Energy for sustainable development in Malaysia: Energy policy and alternative energy. *Energy policy*, *34*(15), 2388-2397.
Mohammadiziazi, R., & Bilec, M. M. (2020). Application of machine learning for predicting building energy use at different temporal and spatial resolution under climate change in USA. *Buildings*, *10*(8), 139.
Moon, J. W., & Kim, J.-J. (2010). ANN-based thermal control models for residential buildings. *Building and Environment*, *45*(7), 1612-1625.
Nägele, F., Kasper, T., & Girod, B. (2017). Turning up the heat on obsolete thermostats: A simulation-based comparison of intelligent control approaches for residential heating systems. *Renewable and Sustainable Energy Reviews*, *75*, 1254-1268.
Nejat, P., Jomehzadeh, F., Taheri, M. M., Gohari, M., & Majid, M. Z. A. (2015). A global review of energy consumption, CO2 emissions and policy in the residential sector (with an overview of the top ten CO2 emitting countries). *Renewable and sustainable energy reviews*, *43*, 843-862.
Nia, E. M., Qian, Q., & Visscher, H. (2022). An Investigation of Occupants' Energy Perceptions in Energy Efficient Retrofitted Residential Buildings: A Review Paper. IOP Conference Series: Earth and Environmental Science,
Nicol, F., & Humphreys, M. (2010). Derivation of the adaptive equations for thermal comfort in free-running buildings in European standard EN15251. *Building and environment*, *45*(1), 11-17.

Nie, P., Roccotelli, M., Fanti, M. P., Ming, Z., & Li, Z. (2021). Prediction of home energy consumption based on gradient boosting regression tree. *Energy Reports*, *7*, 1246-1255.
Okujeni, A., Van der Linden, S., Jakimow, B., Rabe, A., Verrelst, J., & Hostert, P. (2014). A comparison of advanced regression algorithms for quantifying urban land cover. *Remote Sensing*, *6*(7), 6324-6346.
Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022a). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering*, *45*, 103406.
Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022b). Machine learning for energy performance prediction at the design stage of buildings. *Energy for Sustainable Development*, *66*, 12-25.
Paciuk, M. T. (1989). *The role of personal control of the environment in thermal comfort and satisfaction at the workplace*. The University of Wisconsin-Milwaukee.
Padakandla, S., KJ, P., & Bhatnagar, S. (2020). Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, *50*, 3590-3606.
Pallonetto, F., De Rosa, M., D'Ettorre, F., & Finn, D. P. (2020). On the assessment and control optimisation of demand response programs in residential buildings. *Renewable and Sustainable Energy Reviews*, *127*, 109861.
Pan, Y., & Zhang, L. (2020). Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Applied Energy*, *268*, 114965.
Paone, A., & Bacher, J.-P. (2018). The impact of building occupant behavior on energy efficiency and methods to influence it: A review of the state of the art. *Energies*, *11*(4), 953.
Parhizkar, T., Rafieipour, E., & Parhizkar, A. (2021). Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction. *Journal of Cleaner Production*, *279*, 123866.
Parys, W., Saelens, D., & Hens, H. (2011). Coupling of dynamic building simulation with stochastic modelling of occupant behaviour in offices–a review-based integrated methodology. *Journal of Building Performance Simulation*, *4*(4), 339-358.
Paudel, S., Nguyen, P. H., Kling, W. L., Elmitri, M., Lacarriere, B., & Corre, O. L. (2015). Support vector machine in prediction of building energy demand using pseudo dynamic approach. *arXiv preprint arXiv:1507.05019*.
Peng, J., Kimmig, A., Wang, J., Liu, X., Niu, Z., & Ovtcharova, J. (2021). Dual-stage attention-based long-short-term memory neural networks for energy demand prediction. *Energy and Buildings*, *249*, 111211.
Perera, D. W. U., Pfeiffer, C., & Skeie, N.-O. (2014). Control of temperature and energy consumption in buildings-A review. *International Journal of Energy & Environment*, *5*(4).
Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. *Energy and Buildings*, *40*(3), 394-398. https://doi.org/https://doi.org/10.1016/j.enbuild.2007.03.007
Pham, A.-D., Ngo, N.-T., Truong, T. T. H., Huynh, N.-T., & Truong, N.-S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production*, *260*, 121082.

Qureshi, W. A., Nair, N.-K. C., & Farid, M. M. (2011). Impact of energy storage in buildings on electricity demand side management. *Energy conversion and management*, *52*(5), 2110-2120.
Reinhart, C. F., & Wienold, J. (2011). The daylighting dashboard–A simulation-based design analysis for daylit spaces. *Building and environment*, *46*(2), 386-396.
Repository., U. M. L. (2020). *Appliances energy prediction*. https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction
Rollins, S., & Banerjee, N. (2014). Using rule mining to understand appliance energy consumption patterns. Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on,
Sadeghi, S. A., Awalgaonkar, N. M., Karava, P., & Bilionis, I. (2017). A Bayesian modeling approach of human interactions with shading and electric lighting systems in private offices. *Energy and Buildings*, *134*, 185-201.
Santin, O. G. (2011). Behavioural patterns and user profiles related to energy consumption for heating. *Energy and Buildings*, *43*(10), 2662-2672.
Sayed, A. N., Himeur, Y., & Bensaali, F. (2022). Deep and transfer learning for building occupancy detection: A review and comparative analysis. *Engineering Applications of Artificial Intelligence*, *115*, 105254.
Seyedzadeh, S., Rahimian, F. P., Oliver, S., Rodriguez, S., & Glesk, I. (2020). Machine learning modelling for predicting non-domestic buildings energy performance: A model to support deep energy retrofit decision-making. *Applied Energy*, *279*, 115908.
Shabani, A., & Zavalani, O. (2017). Predicting building energy consumption using engineering and data driven approaches: a review. *European Journal of Engineering Research and Science*, *2*(5), 44-49.
Shi, Z., Qian, H., Zheng, X., Lv, Z., Li, Y., Liu, L., & Nielsen, P. V. (2018). Seasonal variation of window opening behaviors in two naturally ventilated hospital wards. *Building and Environment*, *130*, 85-93.
Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. Computing for Sustainable Global Development (INDIACom), 3rd International Conference,
Solomon, D., Winter, R., Boulanger, A., Anderson, R., & Wu, L. (2011). Forecasting energy demand in large commercial buildings using support vector machine regression. *Department of Computer Science, Columbia University, Tech. Rep. CUCS-040-11*.
Somu, N., MR, G. R., & Ramamritham, K. (2020). A hybrid model for building energy consumption forecasting using long short term memory networks. *Applied Energy*, *261*, 114131.
Song, M., Xie, Q., Shahbaz, M., & Yao, X. (2023). Economic growth and security from the perspective of natural resource assets. *Resources Policy*, *80*, 103153.
Stazi, F., Naspi, F., & D'Orazio, M. (2017). Modelling window status in school classrooms. Results from a case study in Italy. *Building and Environment*, *111*, 24-32.
Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and sustainable energy reviews*, *13*(8), 1819-1835.
Syed, D., Abu-Rub, H., Ghrayeb, A., & Refaat, S. S. (2021). Household-level energy forecasting in smart buildings using a novel hybrid deep learning model. *IEEE Access*, *9*, 33498-33511.

Tam, V., Almeida, L., & Le, K. (2018). Energy-Related Occupant Behaviour and Its Implications in Energy Use: A Chronological Review. *Sustainability*, *10*(8), 2635.

Tanimoto, J., Hagishima, A., & Sagara, H. (2008). A methodology for peak energy requirement considering actual variation of occupants' behavior schedules. *Building and Environment*, *43*(4), 610-619.

Truong, L. H. M., Chow, K. H. K., Luevisadpaibul, R., Thirunavukkarasu, G. S., Seyedmahmoudian, M., Horan, B., Mekhilef, S., & Stojcevski, A. (2021). Accurate prediction of hourly energy consumption in a residential building based on the occupancy rate using machine learning approaches. *Applied Sciences*, *11*(5), 2229.

Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761-1768. https://doi.org/http://dx.doi.org/10.1016/j.energy.2006.11.010

Valentina, F., Andersen, R. V., & Corgnati, S. P. (2012). Window opening behaviour: simulations of occupant behaviour in residential buildings using models based on a field survey. 7th Windsor Conference,

Wang, L., Kubichek, R., & Zhou, X. (2018). Adaptive learning based data-driven models for predicting hourly building energy use. *Energy and Buildings*, *159*, 454-461.

Wang, Q., Wong, T.-J., & Xia, L. (2008). State ownership, the institutional environment, and auditor choice: Evidence from China. *Journal of accounting and economics*, *46*(1), 112-134.

Wang, R., Lu, S., & Li, Q. (2019). Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustainable Cities and Society*, *49*, 101623.

Wang, Z., & Ding, Y. (2015). An occupant-based energy consumption prediction model for office equipment. *Energy and Buildings*, *109*, 12-22.

Wang, Z., Hong, T., & Piette, M. A. (2020). Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*, *263*, 114683.

Wang, Z., & Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, *75*, 796-808. https://doi.org/http://dx.doi.org/10.1016/j.rser.2016.10.079

Wei, S., Xu, C., Pan, S., Su, J., Wang, Y., Luo, X., Hassan, T. M., Firth, S., Fouchal, F., & Jones, R. (2015). Analysis of factors influencing the modelling of occupant window opening behaviour in an office building in Beijing, China.

Wen, L., Zhou, K., & Yang, S. (2020). Load demand forecasting of residential buildings using a deep learning model. *Electric Power Systems Research*, *179*, 106073.

Wetter, M. (2011). Co-simulation of building energy and control systems with the Building Controls Virtual Test Bed. *Journal of Building Performance Simulation*, *4*(3), 185-203.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wohlin, C., & Aurum, A. (2015). Towards a decision-making structure for selecting a research design in empirical software engineering [journal article].

Empirical Software Engineering, *20*(6), 1427-1455. https://doi.org/10.1007/s10664-014-9319-7
Xiong, J., Tzempelikos, A., Bilionis, I., Awalgaonkar, N. M., Lee, S., Konstantzos, I., Sadeghi, S. A., & Karava, P. (2018). Inferring personalized visual satisfaction profiles in daylit offices from comparative preferences using a Bayesian approach. *Building and Environment*, *138*, 74-88.
Yan, B., & Malkawi, A. M. (2013). A Bayesian approach for predicting building cooling and heating consumption. Proceedings of 13th International Building Performance Simulation Association Conference,
Yan, D., O'Brien, W., Hong, T., Feng, X., Gunay, H. B., Tahmasebi, F., & Mahdavi, A. (2015). Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings*, *107*, 264-278.
Yao, J. (2014). Determining the energy performance of manually controlled solar shades: A stochastic model based co-simulation analysis. *Applied Energy*, *127*, 64-80.
Yao, M., & Zhao, B. (2017). Factors affecting occupants' interactions with windows in residential buildings in Beijing, China. *Procedia Engineering*, *205*, 3428-3434.
Yoshino, H., Hong, T., & Nord, N. (2017). IEA EBC Annex 53: Total Energy Use in Buildings – Analysis and Evaluation Methods. *Energy and Buildings*, *152*. https://doi.org/10.1016/j.enbuild.2017.07.038
Yu, Z., Fung, B. C., Haghighat, F., Yoshino, H., & Morofsky, E. (2011). A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings*, *43*(6), 1409-1417.
Yu, Z., Haghighat, F., Fung, B. C., & Yoshino, H. (2010). A decision tree method for building energy demand modeling. *Energy and Buildings*, *42*(10), 1637-1646.
Zendeh, E. D. (2019). The Impact of Occupants' Behaviours on Energy Consumption in Multi-Functional Spaces.
Zhang, W., Wu, Y., & Calautit, J. K. (2022). A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal comfort in the built environment. *Renewable and Sustainable Energy Reviews*, *167*, 112704.
Zhang, Y., & Barrett, P. (2012). Factors influencing occupants' blind-control behaviour in a naturally ventilated office building. *Building and Environment*, *54*, 137-147.
Zhao, H.-x., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586-3592.
Zhao, J., Yun, R., Lasternas, B., Wang, H., Lam, K. P., Aziz, A., & Loftness, V. (2013). Occupant behavior and schedule prediction based on office appliance energy consumption data mining. CISBAT 2013 Conference-Clean Technology for Smart Cities and Buildings,
Zhou, X., Yan, D., Hong, T., & Ren, X. (2015). Data analysis and stochastic modeling of lighting energy use in large office buildings in China. *Energy and Buildings*, *86*, 275-287.
Zou, P. X., Xu, X., Sanjayan, J., & Wang, J. (2018). Review of 10 years research on building energy performance gap: life-cycle and stakeholder perspectives. *Energy and Buildings*.
Zuhaib, S., Schmatzberger, S., Volt, J., Toth, Z., Kranzl, L., Eugenio Noronha Maia, I., Verheyen, J., Borragán, G., Monteiro, C. S., Mateus, N., Fragoso, R., & Kwiatkowski, J. (2022). Next-generation energy performance certificates:

End-user needs and expectations. *Energy Policy, 161*, 112723. https://doi.org/https://doi.org/10.1016/j.enpol.2021.112723

Appendix A – Supplementary Material [code] (accessible via hyperlink)

Appendix B – Supplementary Material [simulation input] (accessible via hyperlink)