

# Machine Learning Applications for Energy Utilization of Smart Buildings

Matti Huotari



# Machine Learning Applications for Energy Utilization of Smart Buildings

**Matti Huotari**

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall Tietotekniikantalo/ C202/ T1 on 16 December 2022 at 14.

**Aalto University  
School of Science  
Department of Computer Science  
ASIA group**

**Supervising professor**

Professor Kary Främling, Aalto University, Finland; Umeå University, Sweden

**Thesis advisor**

Professor Heikki Ihasalo, Aalto University, Finland

**Preliminary examiners**

Professor Dante Barone, Federal University of Rio Grande do Sul (UFRGS), Brazil

Professor Jukka Nurminen, University of Helsinki, Finland

**Opponents**

Professor Dante Barone, Federal University of Rio Grande do Sul (UFRGS), Brazil

Professor Jukka Nurminen, University of Helsinki, Finland

Aalto University publication series

**DOCTORAL THESES** 186/2022

© 2022 Matti Huotari

ISBN 978-952-64-1058-6 (printed)

ISBN 978-952-64-1059-3 (pdf)

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

<http://urn.fi/URN:ISBN:978-952-64-1059-3>

Unigrafia Oy

Helsinki 2022

Finland

Publication orders (printed book):

[matti.huotari@aalto.fi](mailto:matti.huotari@aalto.fi)



Printed matter  
4041-0619

**Author**

Matti Huotari

**Name of the doctoral thesis**Machine Learning Applications for  
Energy Utilization of Smart Buildings**Publisher** School of Science**Unit** Department of Computer Science**Series** Aalto University publication series DOCTORAL THESES 186/2022**Field of research** Machine learning applications of smart building energy utilization**Manuscript submitted** 16 November 2022**Date of the defence** 16 December 2022**Permission for public defence granted (date)** 15 November 2022**Language** English **Monograph** **Article thesis** **Essay thesis****Abstract**

Energy utilization of smart or intelligent buildings refers to the definition, modeling, and integration of disparate energy elements into coherent energy systems in buildings with the help of artificial intelligence. A core aspect of applications for smart building energy is to address the issues of energy utilization directly while simultaneously taking into account user-comfort, security and malfunctions. Being deployed in increasing numbers in the built environment, these applications are important components of the built environment today.

Given the risen number of renewable energy sources together with tightened regulation to energy consumption, the smart building energy applications provide means to combine new technology components together with heterogeneous requirements and goals for energy utilization in buildings. These goals comprise of, for instance, optimal scheduling of energy consumption and production, optimization of costs, integration of renewable energy, user-behavior recognition, and consumer comfort.

This research investigates smart building energy applications. This objective is pursued through four research questions which highlight the various aspects of the smart building energy applications: (i) What algorithm to utilize for forecasting the equipment degradation, and what kind of uncertainty is associated with these forecasts for battery packs? (ii) How to build a model in case of gaps or a limited number of observations of interest in data for an air handling unit? (iii) How to involve people in personal environmental comfort decisions for smart building energy applications?, and, finally, (iv) what kind of need is there for smart building energy applications, and which solutions meet these needs? Each of these issues is dealt with using novel techniques, and a related taxonomy was created, as presented in the research publications. The relevance of the proposed solutions is verified with case-studies. Overall, machine learning models solve heterogeneous problems in the field of smart building energy utilization. The results indicate that the proposed solutions can provide answers to a variety of issues regarding building energy management, smart grid, personalization, and maintenance and security.

**Keywords** Machine learning applications for energy utilization, smart buildings, energy utilization**ISBN (printed)** 978-952-64-1058-6**ISBN (pdf)** 978-952-64-1059-3**ISSN (printed)** 1799-4934**ISSN (pdf)** 1799-4942**Location of publisher** Helsinki**Location of printing** Helsinki **Year** 2022**Pages** 170**urn** <http://urn.fi/URN:ISBN:978-952-64-1059-3>



**Tekijä**

Matti Huotari

**Väitöskirjan nimi**

Älyrakennusten energiankäytön koneoppimismalleista

**Julkaisija** Perustieteiden korkeakoulu**Yksikkö** Tietotekniikan laitos**Sarja** Aalto University publication series DOCTORAL THESES 186/2022**Tutkimusala** Älyrakennusten energiankäytön koenoppimismalleista**Käsikirjoituksen pvm** 16.11.2022**Väitöspäivä** 16.12.2022**Väittelyluvan myöntämispäivä** 15.11.2022**Kieli** Englanti **Monografia** **Artikkeliväitöskirja** **Esseeväitöskirja****Tiivistelmä**

Älyenergiaratkaisut tarjoavat joustavaa ja kestäväää energiaa rakennuksissa ja liikenteessä.

Älyenergiasovelluksia käytetäänkin enenevässä määrin rakennetuissa ympäristöissä.

Samanaikaisesti ratkottavia ongelmia älyenergiasovelluksissa ovat energiansäästö, lämpöviihtyvyys, ja epänormaalien tilanteiden hallinta. Uusiutuvien energianlähteiden lisääntynyt käyttö ja lainsäädännön tiukentuneet vaatimukset energiankäytöstä vaativat uusia ratkaisuja, joita älyenergiasovellukset voivat tarjota. Ne soveltuvat erityisen hyvin, jos vaatimuksia on monia tai ne ovat jopa näennäisesti ristiriitaisia. Tällaisia vaatimuksia ovat muun muassa järjestelmien käytön optimointi, kulujen vähentäminen, uusiutuvan energian liittäminen järjestelmään ja käyttäjien lämpöviihtyvyys.

Tämä tutkimus pohtii älyenergiasovelluksia rakennuksissa. Aihetta lähestytään tutkimuskysymysten kautta: (i) Minkä algoritmien avulla voidaan tutkia akkulaitteiston ikääntymistä ja mikä on algoritmista mallista saatujen tulosten epävarmuus? (ii) Kuinka mallintaa, kun pohjadata on epätasapainoissa ja ei ole täydellisen kattavaa (esimerkkinä ilmanvaihtolaitteen data)? (iii) Miten ottaa huomioon ihmiset lämpöviihtyvyyden määrittämisessä ja siihen liittyvässä päätöksenteossa? (iv) Minkälaisia tarpeita on olemassa älyenergiasovelluksille ja mitkä sovellukset tarjoavat ratkaisuja niille?

Jokaista tutkimuskysymystä käsiteltiin uusilla keinoilla, kehitetyt ratkaisut luokiteltiin taksonomisesti, ja kaikki tutkimustulokset julkaistiin lopulta viitenä julkaisuna. Ehdotettujen sovellusten sopivuus varmennettiin tapaustutkimusten avulla. Koneoppimista voidaan soveltaa monitahoisiin ongelmiin rakennusten energiankäyttöön liittyen. Saadut tutkimustulokset antavat viitettä siitä, että ehdotetut sovellukset voivat ratkaista ongelmia, jotka liittyvät energiatehokkuuteen, lämpöviihtyvyyteen ja huoltoon.

**Avainsanat** Koneoppiminen, energiankäytön sovellukset, älyrakennukset, energiankäyttö**ISBN (painettu)** 978-952-64-1058-6**ISBN (pdf)** 978-952-64-1059-3**ISSN (painettu)** 1799-4934**ISSN (pdf)** 1799-4942**Julkaisupaikka** Helsinki**Painopaikka** Helsinki**Vuosi** 2022**Sivumäärä** 170**urn** <http://urn.fi/URN:ISBN:978-952-64-1059-3>



# Preface

Tietäjä löysi tien ja pystyy opastamaan tielle.  
(Kotuksen sananselityksiä: tie)

This dissertation is a result of work in the group of Adaptive Systems of Intelligent Agents (ASIA) at the Department of Computer Science at Aalto University, Finland. This group has offered me an opportunity to be on the road, with lots of emotional events and sheer work both together with people around me and on my own.

First of all, I would like to express my gratitude to my supervisor, Professor Kary Främling. I thank him for his excellent guidance, interesting discussions, and enduring patience throughout my PhD studies. It has been an enjoyable opportunity to be a member of his ASIA research group at Aalto University. Kary also kindly offered me useful data for the research and secured me the freedom to find my own approach to the topics.

I would also like to thank my thesis advisor, Professor Heikki Ihasalo. Through his courses and seminars and many in depth discussions he has directed me on the road of building automation, and he has been able to help construct various testing environments needed in the research. Also, Professor Emeritus Martti Mäntylä and Dr. Avleen Malhi, have both offered their critical help in skillfully editing my sometimes meandering line-of-thought. A special thanks goes to Virpi Kantonen, who rose to the occasion and checked the language of this dissertation at a short notice. The personal interaction with the other ASIA doctoral students, Narges Yousefnezhad, Manik Madhikermi, and Asad Javed, have been few and far apart due to Covid-19 restrictions, but enjoyable, nevertheless. A special thanks for excellent technical help goes to Tuomas Keyryläinen.

The two pre-examiners of this dissertation, Professor Jukka Nurminen and Professor Dante Barone, also deserve a huge thank you for their valuable feedback and suggestions. I am happy to have both of them as the opponents. K.V. Lindholms stiftelse partially funded my PhD studies, which I am grateful for.

On a more personal note, I thank my dear spouse, Anssi Reponen, for supporting me and loving me both through this process and in my personal life. These past years have been a time for letting go of many important people: my best friend, painter Janne Laurila, study comrade Pekka Isto, as well as my father and mother who all encouraged me in my studies. I am grateful for having had them in my life. A big thank you to Dr. Riikka Länsisalmi and Dr. Kimmo Vuorinen, for the many intensive, thought-provoking interactions that have helped keep me grounded in the life outside of my own research topic.

Finally, I am pleased that this road has paid off not only in form of interesting research but also in professional development.

Helsinki, December 2022,

Matti Huotari

# Contents

<b>Preface</b>	<b>1</b>
<b>Contents</b>	<b>3</b>
<b>List of Publications</b>	<b>5</b>
<b>Author's contributions</b>	<b>7</b>
<b>List of Figures</b>	<b>9</b>
<b>List of Tables</b>	<b>11</b>
<b>Abbreviations</b>	<b>13</b>
<b>1. Introduction</b>	<b>15</b>
1.1 Research objectives . . . . .	16
1.2 Research questions . . . . .	17
1.3 Research methodology . . . . .	19
1.4 Research contribution . . . . .	20
1.5 Structure of the dissertation . . . . .	22
<b>2. Theoretical background</b>	<b>23</b>
2.1 Outline of applications for smart building energy utilization	23
2.2 Collecting and transporting the data . . . . .	24
2.2.1 Sensors and IoT devices . . . . .	24
2.2.2 Connectivity and data model . . . . .	24
2.2.3 Data management . . . . .	25
2.3 Equipment in buildings . . . . .	25
2.4 Digital twins . . . . .	26
2.5 Utilizing the data . . . . .	26
<b>3. Articles I-II: time series analysis and regression for state-of-health prediction</b>	<b>27</b>
3.1 Background . . . . .	27

3.2	Related work . . . . .	27
3.3	Case-study methods . . . . .	28
3.4	Case-study implementation . . . . .	31
3.5	Discussion of walk-forward algorithm and state-of-health case-study . . . . .	31
<b>4.</b>	<b>Article III: event detection in imbalanced data</b>	<b>35</b>
4.1	Background . . . . .	35
4.2	Case-study methods . . . . .	36
4.3	Case-study implementation . . . . .	37
4.4	Discussion of undersampling algorithm and event detection case-study . . . . .	38
<b>5.</b>	<b>Article IV: thermal preference classification</b>	<b>39</b>
5.1	Background . . . . .	39
5.2	Related work . . . . .	40
5.3	Case-study methods . . . . .	40
5.4	Case-study implementation . . . . .	40
5.5	Case-study discussion . . . . .	41
<b>6.</b>	<b>Article V: applications for smart building energy utilization</b>	<b>43</b>
6.1	Background . . . . .	43
6.2	Related work . . . . .	44
6.3	Taxonomy for smart building energy utilization . . . . .	45
<b>7.</b>	<b>Conclusions</b>	<b>47</b>
7.1	Implications and limitations . . . . .	48
7.2	Future directions . . . . .	49
	<b>References</b>	<b>51</b>
	<b>Publications</b>	<b>59</b>

# List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

**I** Huotari, Matti and Arora, Shashank and Malhi, Avleen and Främ-  
ling, Kary. A dynamic state-of-health forecasting model for electric  
trucks: Li-ion batteries case-study. In *International Mechanical Engi-  
neering Congress and Exposition IMECE2020*, virtual conference, 84560,  
V008T08A021, 2020.

**II** Huotari, Matti and Arora, Shashank and Malhi, Avleen and Främ-  
ling, Kary. Comparing seven methods for state-of-health time series prediction  
for the lithium-ion battery packs of forklifts. *Applied Soft Computing*,  
111, 107670, 2021.

**III** Huotari, Matti and Främ-  
ling, Kary. Event classification with imbal-  
anced and missing data for an air-handling unit. In *2022 5th Interna-  
tional Conference on Big Data and Artificial Intelligence*, Fuzhou, China,  
5, 82-86, 2022.

**IV** Huotari, Matti and Keyriläinen, Tuomas and Främ-  
ling, Kary. Multi-  
class estimation of human thermal preference for building controls based  
on user feedback and multi-sensor measurements. *Applied Soft Comput-  
ing*, (submitted), 2022.

**V** Huotari, Matti and Malhi, Avleen and Främ-  
ling, Kary. Machine Learn-  
ing Applications for Smart Building Energy Utilization - A Survey.  
*Archives of Computational Methods in Engineering*, (submitted), 2022.



# Author's contributions

## **Publication I: “A dynamic state-of-health forecasting model for electric trucks: Li-ion batteries case-study”**

The author of this dissertation is the primary author of this conference paper. He independently designed the proposed analysis and machine learning methods, implemented the algorithms, and performed the result evaluation. Shashank Arora helped in dissecting the lithium-ion battery data and provided his expertise in the issues of battery physics. Avleen Malhi and other co-authors also contributed to the reviewing of the written text and provided valuable suggestions for the implementation.

## **Publication II: “Comparing seven methods for state-of-health time series prediction for the lithium-ion battery packs of forklifts”**

The author of this dissertation is the primary author of this journal paper. He independently designed the proposed analysis and machine learning methods, implemented the algorithms, and performed the result evaluation. Avleen Malhi contributed to the overall article structuring and methodology selection. The other co-authors provided valuable suggestions for the implementation.

## **Publication III: “Event classification with imbalanced and missing data for an air-handling unit”**

The author of this dissertation is the primary author of this paper. He independently designed the proposed machine learning methods, implemented the algorithms, and performed the result evaluation. Kary Främbling contributed to the experiment setup and provided the 7-year time series data

measured from that environment.

**Publication IV: “Multi-class estimation of human thermal preference for building controls based on user feedback and multi-sensor measurements”**

The author of this dissertation is the primary author of this paper. He independently designed the proposed experiment environment and machine learning methods, implemented the algorithms, and performed the result evaluation. Tuomas Keyriläinen contributed to the laboratory setup and implementing the open message interface to the air handling unit. The other co-authors provided valuable suggestions for the implementation.

**Publication V: “Machine Learning Applications for Smart Building Energy Utilization - A Survey”**

The author of this dissertation is the primary author of this paper. He independently conducted the survey. Avleen Malhi contributed to the structuring and designing the visualization of the article.

**Language check**

The language of my dissertation has been checked by Virpi Kantonen (EFL). I have personally examined and accepted/rejected the results of the language check one by one. This has not affected the scientific content of my dissertation.

# List of Figures

2.1	Some examples of smart building energy applications. . .	24
3.1	The graph of the household power consumption predictions with the extreme gradient boosting model and the corresponding point-wise confidence intervals. Notice that the confidence interval varies due to, e.g., the used aggregation method, the rolling window size and the underlying model's stability. . . . .	32
3.2	State-of-health values for a battery during three-month-period with confidence intervals produced by gradient boosting and walk-forward repetitions. The blue line is the predicted and the orange is the observed SoH. The shadowed area indicated point-wise 95% confidence interval for predictions. . . . .	32
3.3	Battery (b-f) GB predicted SoH values. The blue line is the predicted and the orange is the observed SoH. The shadowed area is +/- two standard deviations. . . . .	33
3.4	All SoH values together with ARIMA (0,1,1) predicted SoH values. . . . .	34
4.1	Recorded events in 2014—2021. . . . .	38
5.1	The overall architecture of the comfort model case-study.	41
6.1	Taxonomy of applications for smart building energy utilization. . . . .	46



# List of Tables

1.1	Summary of the publications and related research questions ( <b>X</b> = fully addressed, <b>x</b> =partially addressed). . . . .	19
5.1	Summary of the experiment subjects for evaluation of personal thermal comfort in an office room. . . . .	41
6.1	Identified key surveys for building applications. . . . .	44



# Abbreviations

**AHU** air handling units

**AI** artificial intelligence

**API** application interface

**ARIMA** autoregressive integrated modeling average

**BAG** bagging with decision tree as the base estimator

**CI** confidence interval

**Covid-19** coronavirus disease of 2019

**EU** the European Union

**GB** gradient boosting

**HTML** HyperText Markup Language

**HTTP** HyperText Transfer Protocol

**HVAC** heating, ventilation, air conditioning and cooling

**IoT** Internet-of-Things

**MAE** mean absolute error

**ML** machine learning

**NMC** nickel–cobalt–manganese

**O-DF** open data format

**O-MI** open messaging interface

**REST** representational state transfer

**RSME** root mean squared error

Abbreviations

**SoH** state-of-health

**XGB** extreme gradient boosting

**UN** the United Nations

# 1. Introduction

Smart or intelligent energy refers to definition, modeling, and integration of disparate elements into coherent energy systems [1]. Moreover, smart building energy applications consist of heterogeneous machine learning models to solve problems related not only to building energy utilization directly but also to user-comfort and abnormal situations [2, 3]. Smart energy has emerged to provide flexible and sustainable energy for buildings and transportation.

Given the rising number of renewable energy sources together with legal restraints to energy consumption, smart building energy applications provide means to combine the heterogeneous goals for energy utilization in the context of buildings [4, 5]. These goals comprise of, for instance, optimal scheduling of consumption and production, optimization of costs, integration of renewable energy, utilization of energy storages, user-behavior recognition, and consumer comfort.

This dissertation explores and implements several statistical analysis and machine learning methods with case-studies for research of smart building energy applications. The focus of this dissertation is to demonstrate with case-studies the pertinence of the machine learning models to heterogeneous problems in the field of smart building energy utilization.

Due to the dominant role of energy in everyday life, the importance of smart building energy applications is increasing. However, the applications target not only to reduce energy costs and to decrease energy consumption but also to increase well-being of their users. To achieve these goals, several of them need to be combined in an application. Moreover, the integration of an application requires inclusion of an operational control, electric devices, renewable energy sources, and an energy storage system, or a subset of these. In addition, to create an application that has energy utilization and user-comfort in balance, comfort needs to be evaluated and integrated into the application. Finally, there may be some particular needs, for instance, to be involved personally in decisions how to utilize energy in a building. Another example of a particular need is monitoring the energy system for fault situations. The smart building energy applications can be regarded

as a part of smart cities, which is an ecosystem for applications, buildings, cities, and service providers to interact with each other. Such an ecosystem combines many applications to meet the needs of their user in governed manner.

Although building energy management, smart grid, personalization, and maintenance and security applications are needed for smart buildings, the implementations face challenges due to the heterogeneous tools and environments. The absence of applicable data is perhaps the primary issue when making these applications. This includes both the absence of public databases to make models, and the limited availability of appropriate sensor data from buildings. The absence of well-defined personal data is another issue when making applications due to perceived difficulty of measuring user preferences and behavior. Furthermore, data collected may have gaps and be imbalanced. This dissertation addresses these challenges through selected case-studies of applications in buildings by giving special attention to data utilized. As a result, the applicability of various applications in the context of smart buildings is the main theme of this study that refines the previous research on smart cities [6].

## 1.1 Research objectives

The overall objective of this research is to create smart energy applications for buildings that generalize well to different buildings and environments. In particular, the applications should reach adequate quality measures, such as high-enough prediction accuracy. This research examines the development of selected applications with several machine learning methods, and then compares the suitability of the selected methods to the problem in question including considerations of data and quality measures. In order to create such smart energy applications, the challenges hindering the wider adaptation of smart energy applications in buildings need to be addressed. To solve the challenges motivating this research, we aim at the following targets:

- Evaluation of degradation of equipment over time. To offer flexibility for the energy production and consumption, batteries can store energy to be used on demand. The physical phenomena related to this type of mechanical equipment can cause usage limitations that need to be addressed. Typically, this is done with evaluation of the level of measure of degradation, such as state-of-health of batteries.
- Creation of machine learning model in case of imperfect data recordings. To overcome the absence of large scale labeled data to create a model, some means to make more labeled data are needed. One way to overcome

the problem of lacking data is to make use of all available data to the extent possible. To create a model and to predict with the data collected in presence of missing data records and a limited number of observations of interest, developing a model that overcomes these limitations is one means to achieve that.

- Involvement of people in comfort decisions. A balance between energy-efficiency and user-comfort is needed to make the smart building energy applications widely adapted. Traditionally a median of past opinions is utilized for smart building energy application modeling, for instance for room temperature setting. However, a straight-forward way to determine the personal comfort is to inquire about it. How to combine the direct inquiry into an energy control model is an emerging topic with few realized applications.
- Analysis of the elements of smart building energy and systematic classification of the disparate elements. A systematic approach to map the needs clarifies the gaps in the present applications. Moreover, the relationship between the different domains are sometimes overlapping and sometimes far apart. A systematic approach to classify these applications also helps to understand these relationships. This helps the integration of disparate models that is essential to fully exploit the models in buildings.

## 1.2 Research questions

This dissertation focuses on the smart building energy with experimental case-studies of applications. To give an overview, it presents a novel taxonomy for these applications to put the case-studies into relevant context. In particular, it answers the following four research questions:

***Research question 1:*** *What algorithm to utilize for predicting the equipment degradation and what kind of uncertainty is associated with these predictions for battery packs?*

Time series methods, such as autoregressive integrated modeling average (ARIMA) based statistical techniques [7] or machine learning techniques have traditionally been utilized to yield estimations on the degradation of cells of batteries. Publication I employs battery packs that combine several cells both in series and in parallel. This requires rigorous degradation forecasts as heat generation in battery packs increases as they age. If unchecked, it can cause internal short circuits in these relatively powerful pieces of equipment compromising the safety of buildings and their users.

Publication I presents the development and evaluation of a model for one battery pack and, to ensure the generalization of the model, set of similar battery packs. Publication II presents further developed results. The experiment had a greater number of battery packs for its use and it deepened the research on degradation by comparing seven models to it.

***Research question 2:*** *How to build a model in case of gaps or a limited number of observations of interest in data for an air handling unit?*

Reliably fault detection in air handling units (AHU), or other building equipment, is a key aspect of correcting faults and eliminating non-optimal functionality. Machine learning classification methods with data resampling are widely applied to detect these kinds of events, which, by their nature, seldom occur in equipment. Some events are relatively rare, such as faults in air handling units or other building equipment. Therefore, for creating models that yield more accurate and precise [8] results can sometimes be achieved by combining resampling methods with other machine learning methods. Publication III compares several resampling methods and develops an oversampling method to mitigate the problem of an imbalanced data set with some missing data.

***Research question 3:*** *How to involve people in environmental comfort decisions in an office room for smart building energy applications?*

Personal opinions and feelings are difficult to evaluate based on predefined median value or physical measurements only. However, a predefined median opinion, or some more refined method is needed to set environmental values, such as room temperature. As modern buildings tend to have similar designs and react alike all over the world, there is ample room to develop machine learning models to make informed decisions about personal environments that could be applicable in many buildings. Publication IV discusses the factors needed to determine the personal thermal comfort for building controls in an office room.

***Research question 4:*** *What kind of need is there for smart building energy applications and what solutions meet these needs for smart building machine learning application?*

The application areas of smart building energy utilization consist of building energy management, smart grid, personalization, and maintenance and security. The old techniques to monitor, analyze and control are succumbing to more flexible, if not to say, more intelligent machine learning techniques. This is already noticeable in cases where new technologies, such as solar panels, are introduced to buildings. Furthermore, the current

**Table 1.1.** Summary of the publications and related research questions (**X**= fully addressed, **x**=partially addressed).

	Research Questions	Publications				
		I	II	III	IV	V
<b>RQ1</b>	What algorithm to utilize for forecasting the equipment degradation and what kind of uncertainty is associated with these forecasts?	<b>X</b>	<b>X</b>			<b>x</b>
<b>RQ2</b>	How to build a model in case of gaps or few observations of interest in the observations?	<b>x</b>	<b>x</b>	<b>X</b>	<b>x</b>	<b>x</b>
<b>RQ3</b>	How to involve people in environmental comfort decisions in an office room for smart building energy applications?				<b>X</b>	<b>x</b>
<b>RQ4</b>	What kind of need is there for smart building energy applications and what solutions meet these needs?	<b>x</b>	<b>x</b>	<b>x</b>	<b>x</b>	<b>X</b>

energy embargoes have risen the need for allocating the existing resources more minutely. Publication V surveys the current research in this field. It classifies the applications of smart building energy utilization into a novel taxonomy based on the survey results.

Table 1.1 presents the research questions addressed in the respective publications. In this table, capital X refers to the related Research Question (RQ) that is fully addressed and lowercase x is for the partially addressed RQ.

### 1.3 Research methodology

Machine learning is concerned with automatically detecting patterns in data, and then use these disclosed patterns to forecast future data, or performing other kinds of decision making under uncertainty. This dissertation harnesses mostly quantitative but also some qualitative methods for the research on machine learning algorithms. Each of the methods are described in more detail with the respective case-study, as several case-study specific methods were employed.

The initial scope of the work was identified during the literature review and interviews with the field experts of building automation. For each case-study, an initial hypothesis and research questions were formed at this stage. This dissertation systematically employs controlled experiments to compare different machine learning methods on the data [9]. Moreover, a resampling algorithm was developed based on previous theoretical work on resampling algorithms.

For the controlled experiments of this work, data collection was followed by model development and model evaluation. To frame a problem more explicitly, exploratory data analysis including summarization and visualization was adapted. Then a significant part of this dissertation concerns

methods to select, clean and prepare the data for further model development, particularly as in many cases the observations had no labels attached to classify or quantify them. That is, to make these labels available for machine learning purposes, in some case-studies basic physics was applied to the data, in others an application to inquire about personal feelings and qualities was applied. In one case-study, a resampling algorithm was developed for data pre-processing purposes.

To quantify the uncertainty of the models, estimation statistics were used to yield confidence intervals for predictions for some of the models developed.

## 1.4 Research contribution

This dissertation summarizes the author's primary contribution in the form of five peer-reviewed publications, as detailed below.

*Publication I* presents a dynamic state-of-health prognosis model for lithium-ion battery packs. The novel contribution of this publication is that it is one of the first published pieces of research on the state-of-health prognosis on battery-pack level, as we had a unique data set of 31 lithium-ion battery packs from forklifts in commercial operations. Despite data being collected is from electric trucks, the same battery packs have a second life as energy storages in buildings, where the same forecast model can be applied. We proposed an autoregressive integrated modeling average (ARIMA) model to evaluate the state-of-health value to assess the usability of the battery. The results indicate that the developed model provided relevant tools to analyze the data from several batteries. Moreover, we compared it with a machine learning model, bagging with decision tree as the base estimator (BAG). Even though this comparison was very rudimentary its results indicate that the developed supervised learning model using decision trees as base estimator yield reasonable results.

*Publication II* deepens the work in Publication I. In this publication the authors performs a thorough comparison with seven different machine learning models to yield a model for the health-of-state evaluation. This publication contributes to the literature by introducing a novel gradient boosting model for SoH prediction based on real-world application of lithium-ion battery packs in forklifts and implementation of a novel walk-forward algorithm for validating the models. The results about the final model suggest that we were able to enhance the results both for one battery pack and as a set of these battery packs in respect to the previously developed model. Moreover, we further validated the model for extracting cycle counts presented in our previous

work with data from new forklifts; their battery packs completed around 3000 cycles in a 10-year service period, which corresponds to the cycle life of commercial Nickel–Cobalt–Manganese (NMC) cells. Our data was unique, as far as we could confirm, as there are no public databases on battery pack data; therefore, we made ours available at <https://github.com/huotarim/huotarim-xgboost-li-ion-batteries>.

*Publication III* presents solution to a challenge that is not entirely dissimilar to challenges in Publications I and II: how to process the raw data so that it can be utilized for more enhanced model development. Publication III contributes to the field of imbalanced classification problems by proposing a novel data undersampling algorithm to enhance the classification model results in the presence of imbalanced and missing data. Then, the paper proposes a model to forecast imperfect heat recovery events in an air handling unit that occur relatively seldom. We had 7-year data from an air handling unit (AHU) that had imbalanced class distribution between the majority class and the minority class. We wanted to develop a predictive model for the minority (target) class that presented the observations classified as non-optimal functionality. Moreover, the minority class was distributed unevenly and, overall, the data contained missing data sequences. Finally, we compared several resampling methods and machine learning methods in this publication to create the proposed classification model.

*Publication IV* ponders on personal thermal comfort that is a key aspect for enhancing the indoor environment quality and the energy efficiency of buildings. This paper contributes to the field of thermal comfort studies by yielding multi-class estimations for personal thermal preference that rely on the time-dependency of the data. Moreover, this paper compares several combined machine learning methods for extracting the best estimation. The developed model employs variables, such as physical factors, environmental factors, actions of the users and information about the outdoor environment, to make an estimation of personal thermal preference for building controls in an office room. We utilized multiple sensors combined with a simple user survey. The results of the experiment suggest that utilizing measurements from a user survey, a wristlet, office room sensors, and a nearby weather station, predicted the resulting thermal preference class satisfactorily. In the experiment, the model generalized reasonably to different people in the same office room environment and wearing similar clothing.

*Publication V* presents a systematic review of the literature for smart building energy applications and then creates a novel taxonomy for these. The contributions of this paper are a comprehensive review of the smart

building energy applications and creation of a taxonomy for these solutions containing building energy management, smart grid, personalization, and maintenance and security solutions. This paper puts into a context the applications and solutions developed in Publications I-IV.

## **1.5 Structure of the dissertation**

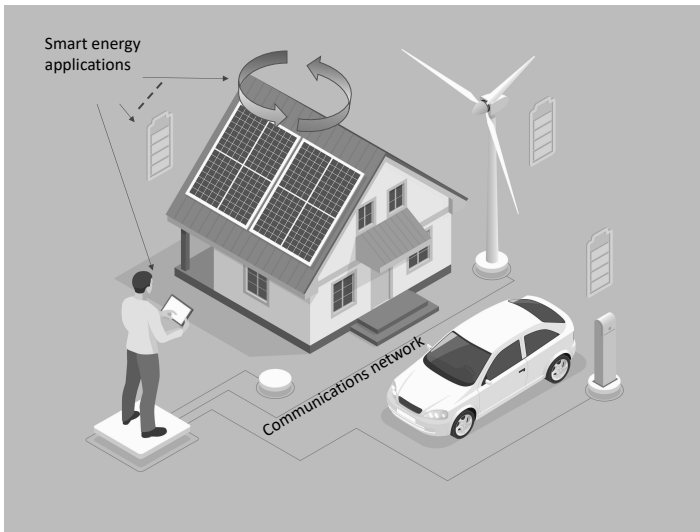
This dissertation is further structured into five chapters. Chapter 2 describes the key technological foundations that are required to understand this dissertation, particularly the data processing and machine learning computing technologies needed to develop applications for smart building energy utilization. Chapter 3 provides insights into the main research findings presented in Publications I and II. This chapter primarily discusses time series and solutions for firmly time-dependent elements of buildings. It also presents a novel walk-forward algorithm to lessen the computational burden to some extent. Chapter 4 presents a resampling solution to encounter bias in data. Chapter 5 presents a thermal comfort assessment model for building controls in an office room. Chapter 6 summarizes the smart building energy applications that cover the domains of building energy management, smart grid, personalization, and maintenance and security. The research findings listed in this chapter are detailed in mostly in Publication V, but also in Publications I-IV. Finally, this dissertation is concluded in Chapter 7 along with the future research directions.

## 2. Theoretical background

This chapter describes the technologies material to the problem domain of this dissertation. The technological understanding provides the appropriate tools to build solutions in the area of smart buildings. Therefore, a good understanding of related tools is required to achieve the research objective of this dissertation. This chapter begins by outlining a classification of smart building energy applications from the perspective of collecting and utilizing data. Section 2.1 makes an overview of smart building energy applications. Section 2.2 introduces various data collection and transport techniques related to it. The Section 2.5 discusses the common goal for all of the smart energy applications together with some machine learning model implementation aspects.

### 2.1 Outline of applications for smart building energy utilization

Smart building energy applications extends from applications dealing with building energy utilization directly to user-comfort, personalization, maintenance and security [10, 11, 12]. In recent years, several machine learning methods have emerged that, together with advances in the sensor and Internet-of-Things (IoT), create a well-founded basis for making smart building energy applications. This section introduces some of the relevant concepts, technologies and building equipment linked to the topic of this dissertation. Figure 2.1 depicts a building listing some equipment and people in the context of this dissertation. They include battery-packs to store energy, an air handling unit, building automation (heating, cooling, ventilation, air handling units), persons living in the house and, finally, the smart building energy applications. In a wider context, they contain renewable energy production, other electric appliances, connections to wider environment (utility grid), and security and maintenance that this dissertation presents later.



**Figure 2.1.** Some examples of smart building energy applications.

## 2.2 Collecting and transporting the data

### 2.2.1 Sensors and IoT devices

Several types of sensors, devices, other information systems, as well as people, can be connected to collecting a large amount of real-time data. Some data utilized can be historical or represent a mean theoretical value of a variable. The majority of the data sources are low-cost devices, often integrated into, for example, an air handling unit. The rising number of IoT equipment and low cost of sensors have created a suitable environment for wider utilization of smart building energy applications. This has happened in conjunction with other developments, such as availability of computational power in the cloud, adaptation of heat pumps, and installation of renewable energy equipment in houses and apartments.

### 2.2.2 Connectivity and data model

There are several connectivity possibilities of the devices, such as WiFi, that is the ubiquitous IEEE 802.11 standard based local area network technology. Depending on the related implementation, other elements of communication and data technologies are needed, such as data transport protocol and some data management technology in field operations. This is a non-trivial domain to consider when building an application, especially with numerous building automation communication protocols in the field today. For instance, a typical interface in the EU is REST API [13] that specifies the data transfer between an appliance and a machine

learning algorithm. However, it has a response time that may not satisfy requirements for minute energy trading, and much effort is needed by the model developer to access the often proprietary lower level communication layers and interfaces. Experiments in this dissertation, constructed by ourselves, adopt the open communication standards, O-MI and O-DF [14, 15], defined by the Open Group consortium. These standards provide real-time interaction between the sensors in equipment and services used for saving the data, similar to HTML/HTTP protocols for the web, enabling any equipment with an interface to be connected [16, 17].

### **2.2.3 Data management**

The number of buildings with sensors containing equipment is rising. The traditional three dimensional building models and ontologies respond poorly to time temporal aspects of the sensor-collected data from the buildings nowadays. Therefore it is of utmost importance to pay attention to managing a large amount of real-time or nearly real-time data derived from buildings into the smart energy applications.

## **2.3 Equipment in buildings**

Smart energy applications combine mechanical equipment, sometimes even old technology in the solutions. Chemical batteries are an example of ancient technology still in use. Modeling it with machine learning models is in principle easy: a battery is easy to equip with sensors. However, the difficulty lies in the plethora of chemistries and nominal powers of the batteries in the field. The machine learning models tend to have a limited adaptability if one of these parameters change. Nevertheless, batteries are used as energy storage in smart solutions.

A lot of energy consumption is related to heating, ventilation, air conditioning and cooling (HVAC). The realizations in this field are many. A simple ventilation system may control the air flow and quality. This is not an air conditioning system, but is still utilized as an adequate solution in many buildings. Air conditioning system adjusts the temperature indoors on top of ventilation. A big building may have one or few HVAC equipment and some air handling units (AHU) in rooms. In cool climates, this enables to circulate the warm indoor air while taking some fresh air from outdoors to keep the air quality satisfactorily in each room. The fresh air may be dried, impurities filtered out, and the temperature is adjusted.

In traditional houses, there is typically a ventilation system. Recently, however, more and more heat pumps are being installed, as have even more advanced HVAC systems. The terminology is getting fuzzy, as "air conditioning" can mean different things to different people. Hence, the type

of equipment, functionality of the equipment, and terminology utilized (ventilation, AHU, HVAC) needs to be clear when developing a model even for experiments not to mention the field conditions. This affects machine learning modeling significantly. In turn, it is important to set the expectation level correctly when a machine learning model is created. What is the expected application field? How tight a similarity for the environment is required, or can be expected for that matter? And finally, does the model need to be applicable in the field operations?

## **2.4 Digital twins**

We utilize the digital twins architectural information system model for implementing smart building energy applications for built environment for applicable parts [18]. The technological challenge for developing smart building energy applications, for, for example, product life-cycle management, is the high number and variety of information systems that need to communicate over organizational limits and over time. We extend the findings on this subject of the paper [19], for example, by making a case-study on lithium-battery degradation over time.

## **2.5 Utilizing the data**

The United Nations (UN) launched sustainable development goals in 2015 that include goals for sustainable energy [4]. Local adaptations, such as Fit for 55 in the European Union suggest a more detailed regulation to save energy and to utilize a wider range of renewable energy sources [20]. The adaptation of the regulation require practical smart building energy applications to meet targets set by regulation. In addition, a clear correlation exists between household energy consumption and user behavior, personal comfort and finances being the outstanding factors for behavior [21]. This dissertation adopts several techniques to meet this overall goal of flexible and sustainable utilization of energy. However, the solutions to meet this goal are heterogeneous and, at the same time, various technology domains can contribute to implement these solutions as the presented case-studies and the survey illustrate.

### **3. Articles I–II: time series analysis and regression for state-of-health prediction**

This chapter sets to resolve application challenges introduced in Chapter 1 that are related to equipment state-of-health. The challenges of the imperfect data recordings and the overall need for this application are included in the reflections of this chapter. The state-of-health challenges are a part of the first research question (RQ1) identified in Publications I and II. The imperfect data challenge is part of the second and fourth research questions (RQ2, RQ4) identified in Publications III and V.

#### **3.1 Background**

Smart building energy applications are required to ensure the safety of their use. In case of battery equipment this is imperative as the battery may self-ignite as it wears out. The goal of this chapter is to investigate an appropriate algorithm to evaluate the state-of-health with enough future time-horizon to improve the safety in terms of evaluating the time to replace the battery used as an energy storage by a solution. Overall, the time-series and uncertainty estimation of the proposed model is also investigated in this chapter.

#### **3.2 Related work**

Utilizing energy storage systems has been investigated in literature, and can be classified according to the energy stored: electric, thermal, or kinetic. Within smart building energy applications the energy storage is practically expected to serve two purposes: the integration of renewable sources and the storage of energy for utilization. The optimal time to utilize energy is not necessarily the time when it is produced [22]. Batteries are promoted to be used as the storage, as they are easily available and the vehicle industry promotes a second-life use for them [23]. The lithium-ion battery packs in storage systems can benefit from the added reliability and safety assurance

by a fast, yet accurate, state-of-health (SoH) prediction. Traditionally, SoH forecasting has relied on equivalent-circuit models; however, more recently statistical and machine-learning techniques have been proposed, including ARIMA based statistical methods and several machine learning methods, such as gradient boosting (GB) [24]. The studies indicate that models can satisfactorily predict the SoH. Nevertheless, it is also known that the relationship between the observed signals and the SoH is complex under real conditions, and the complexity is increased as many cells need to be combined in series and in parallel to build up battery packs that are then used as energy storages [25]. To overcome the problems, in contrast to several recent studies that model a battery pack based on unit cells, we employed a battery data set from lithium-ion battery packs that are used in electric forklifts.

### 3.3 Case-study methods

In the field of machine learning, gradient boosting pertains gradient-based optimization of models composed of consecutive transformations [26, 27]. Gradient boosting has turned out to be considerably successful in areas such as robotics [28, 29], risk assessment [30, 31], and end-of-life estimation [32, 33]. In part, the success of gradient boosting is due to the insertion of freedom into the model design, for example, freedom to the choice of the most appropriate optimization function. However, this also introduces a lot of trial and error to extract that most appropriate function. This does not cause a major problem, as the gradient boosting models are relatively simple to implement.

In the following, I will briefly present the basics of the gradient boosting regression, the model training with walk-forward and empirical mode decomposition, which are an important part of Paper II. I will conclude by introducing ARIMA statistical method that is an important part of Paper I, and provides a traditionally used base-line for the later developed machine learning models.

Gradient boosting regressors are additive models that need inputs  $x_i$ , targets  $y_i$  and estimators  $h_m$  that best fit these data, so that a prediction  $\hat{y}_i$  can be made as:

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^M h_m(x_i) \quad (3.1)$$

The gradient boosting models are built by fitting a new tree:

$$F_m(x) = F_{m-1}(x) + h_m(x), \quad (3.2)$$

The tree is fitted so that the losses are minimized given the previous ensemble of trees:

$$F_m(x) = \arg \min_h L_m = \arg \min_h \sum_{i=1}^n l(y_i, F_{m-1}(x_i) + h(x_i)), \quad (3.3)$$

where  $l$  is a loss function that can be approximated as:

$$l(y_i, F_{m-1}(x_i) + h(x_i)) \approx l(y_i, F_{m-1}(x_i)) + h(x_i) \left[ \frac{\partial l(y_i, F(x_i))}{\partial F(x_i)} \right]_{F=F_{m-1}}. \quad (3.4)$$

As the loss is differentiable for any given  $F_{m-1}(x_i)$ , this can be denoted as:

$$h_m = \arg \min_h L_m = \arg \min_h \sum_{i=1}^n h(x_i) g_i. \quad (3.5)$$

At each iteration the estimator  $h(x_i)$  is fitted to predict a value that is targeted to be proportional to the negative gradient  $-g_i$  of the samples, which minimizes the equation (3.5). A widely used loss function is mean absolute error (MAE) [34] over the samples.

In this case-study we re-framed the multivariate time series as a supervised learning problem to define the number of past time steps used for making a forecast and to define the number of prediction timesteps for the prediction horizon [35]. In the model tuning phase, we split this re-framed data to training data and test data. The walk-forward we implemented for this case-study can proceed one-step-ahead or multi-step-ahead at each iteration round, and can either utilize all historical data (called expanding window approach) or drop the oldest data (called sliding window approach). The algorithm implemented is Algorithm 1.

As this algorithm utilizes the sliding window method, the successive training sets are not super-sets of those coming before them, and this yields models that have more variation; however, at the same time, some of the training data is lost. In addition, by producing results with different sets, the algorithm can yield point-wise confidence intervals (CI) that quantify the uncertainty for the predictions [36].

$$\widehat{SE}_{\text{SoH}}(\hat{\mu}_t(n)) = \frac{\hat{\sigma}_t(n)}{\sqrt{n}} \quad (3.6)$$

where  $n$  are the predictions made by the novel Algorithm 1.

To evaluate the rate of deterioration of a battery pack, the empirical mode decomposition is a mathematical time domain decomposition method, which can convert a group of time series into locally narrow band components, the intrinsic mode functions [37]. This method is applied to, e.g., asserting power quality [38], or predicting remaining useful lifetime of lithium-ion batteries [24]. In addition, the decomposed function can be transformed, and an instantaneous phase can be derived from it. In turn, an instantaneous frequency can be obtained through the derivative of the

**Algorithm 1** Walk-forward: sample predictions with point-wise confidence intervals

---

```

1: inputs:
2:   Obs: time series' observations that are re-framed as supervised learning problem (re-framed to windows)
3:   nS: sample of windows utilized for testing
4:   nroll: number of windows stepped over in a window roll
5: local variables:
6:   Te: a sample of time series' windows
7:   Tr: time series' windows preceding Te
8:   RTe: a rolled sample of time series' windows
9:   Wp: consequent predictors utilized for making predictions
10: outputs:
11:    $\widehat{SoH}$ : SoH predictions for the sample
12:   CI: point-wise confidence interval for these SoH predictions
13:   Wt: consequent targets (ground truth)
14: require:
15:    $|Obs| > 1$ 
16:    $n_S > 0$ 
17:
18:   Tr, Te ← split Obs to train and test sets according to nS
19:   RTe ← [ ]
20:   for R ← 0 to  $|Te| \bmod n_{roll}$  do
21:     RTe ← append window (Te[R])
22:   end for
23:    $\widehat{SoH}, CI$  ← [ ], [ ]
24:   for T ← 0 to  $|RTe|$  do
25:     Wp, Wt ← RTe[T] separate predictors and targets for this iteration step
26:      $\widehat{SoH}$  ← fit the model with Tr and predict with Wp
27:     CI ← append  $\pm 1.98 * SE(\widehat{SoH})$  ▷ Eq. 3.6
28:     Tr ← append windows from Te until and including window RTe[T]
29:     if sliding window then
30:       Tr ← delete  $|n_{roll}|$  windows from head of Tr
31:     end if
32:   end for
33:   return  $\widehat{SoH}, W_t, CI$ 

```

---

instantaneous phase. This latter one is a visually easily explainable and powerful tool to detect changes in frequency of the observed phenomenon.

Finally, in the field of time series analysis methods, ARIMA, or autoregressive (AR), integrated (I), modeling average (MA) is utilized to learn the suitable ARIMA model from the input time series  $Y$  to the estimator  $\hat{Y}$  for create a forecast model. This is a baseline model, to which other machine learning models are often compared. If the time series' statistical properties (e.g. the variance) are not constant over time, the time series is transformed to a stationary one. A simple data transformation is differencing; however, here it must be emphasized that the applicable data transformation method is data-dependent. Differencing is made as follows:

$$y_t = Y_t - Y_{t-1}, \quad (3.7)$$

where  $y$  denotes the differenced time series made stationary. From this, the stationary estimator values,  $\hat{y}_t$ , are calculated by as follows:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \dots + \theta_q e_{t-q}, \quad (3.8)$$

where  $\mu$  is constant,  $\phi_i y_{t-i}; i \in 1, \dots, p$  is a lagged value of time series (AR-terms) that is stationary, and  $\theta_i e_{t-i}; i \in 1, \dots, p$  a lagged error (MA-terms). The forecast is yielded by differencing in reverse manner:

$$\hat{Y}_t = \hat{y}_t - Y_{t-1}. \quad (3.9)$$

The denotation of the yielded model is  $ARIMA(p, d, q)$ , where  $p$  is AR-term,  $d$  is number of differencing done and  $q$  is the MA-term. We utilized a root mean squared error (RMSE) as the evaluation metrics for the model development [39].

### 3.4 Case-study implementation

To evaluate the walk-forward algorithm, we verified it against public dataset: the household power consumption data set [40]; this data was aggregated to monthly values. The model that was used to predict the monthly power consumption was a simple extreme gradient boosting (XGB) [41], where the number of estimators was 50, maximum depth was 2. For the verification, the number of observation windows used for walk-forward was 30, the size of the sliding window was 7 and the number of rolled over windows was 4.

To evaluate the models against the unique lithium-ion battery pack data employed, we executed a basic comparison of several machine learning models with walk-forward method utilizing both of expanding and sliding window approaches to find the best model in terms of MAE metrics. In this paper, we added upper and lower confidence intervals (CI) to each of the point-wise predictions for the final model selected, the gradient boosting, (Figure 3.2).

To evaluate how well the model generalized to other battery pack, we executed a Wilcoxon statistical test to verify, which batteries come from the same distribution, as this is a widely employed verification method to initially screen whether the model is applicable for a piece of data [42].

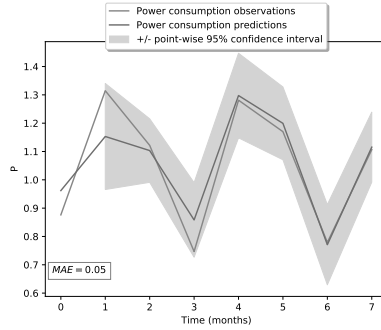
To evaluate the possible change rate of the deterioration of the battery mechanical quality, instantaneous frequency was obtained for the derived state-of-health signal. This latter one is a visually easily explainable and powerful tool to detect changes in frequency of the observed phenomenon as normally the instantaneous frequency tends to follow a straight line calculated from mean values of these instantaneous frequencies in case a constant frequency is observed for a phenomenon.

Finally, the ARIMA provided a solid and well-utilized baseline for our time-series modeling and for further machine learning model development. We assumed that SoH linearly decreases based on an initial data analysis.

### 3.5 Discussion of walk-forward algorithm and state-of-health case-study

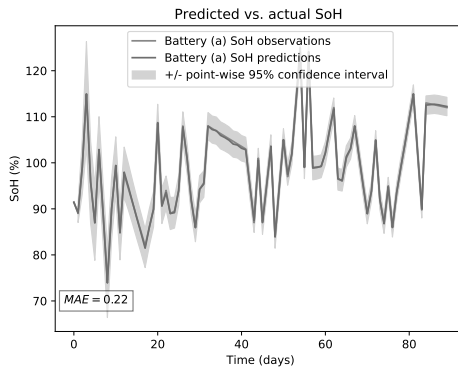
The model for verifying the walk-forward algorithm proposed was predicting the household power consumption from a public database. The

Algorithm 1 yielded the root mean square of error (RMSE) 0.08, which exceeded the naïve model’s RMSE of 0.11 [39]. The yielded prediction results with the corresponding point-wise confidence intervals (CI) are in Figure 3.1. The CI does not fluctuate significantly nor show a clear trend; the model seems to be stable [36].



**Figure 3.1.** The graph of the household power consumption predictions with the extreme gradient boosting model and the corresponding point-wise confidence intervals. Notice that the confidence interval varies due to, e.g., the used aggregation method, the rolling window size and the underlying model’s stability.

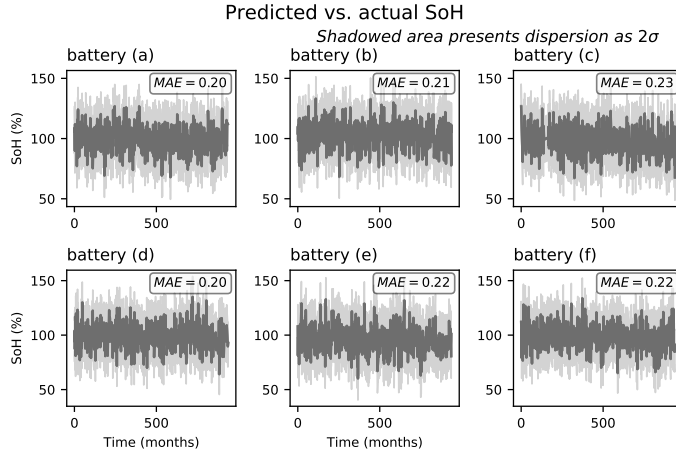
The best model for the battery state-of-health predictions was gradient boosting with sample size 30, window size 14, and expanding window one-step at time. The results are illustrated in Figure 3.2.



**Figure 3.2.** State-of-health values for a battery during three-month-period with confidence intervals produced by gradient boosting and walk-forward repetitions. The blue line is the predicted and the orange is the observed SoH. The shadowed area indicated point-wise 95% confidence interval for predictions.

To apply the model for a set of battery packs, the hypothesis (H0) was set as follows: the sample distributions from different batteries were related to the battery (a). Wilcoxon yielded that 65% of the batteries had the same distribution as battery (a) utilized to develop the model (failed to reject H0), and consequently 35% had different distribution (rejected H0). The model applied to the batteries coming form the same distribution yielded

reasonable results.



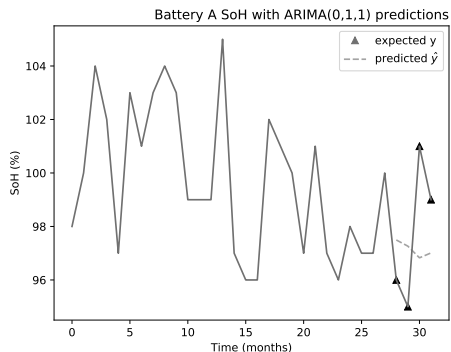
**Figure 3.3.** Battery (b–f) GB predicted SoH values. The blue line is the predicted and the orange is the observed SoH. The shadowed area is  $\pm$  two standard deviations.

Other results on sliding window versus expanding window on the same data indicate that the expanding window method yields, in majority of the cases, slightly better MAE and RMSE values for this data set than the sliding window method. This is in accordance with the general findings in industry in data with relatively few samples [43]. Furthermore, the same results indicate that a window roll that is up to 23–28% of the utilized window size is applicable to this data set for yielding reliable results.

To detect the changes in the battery pack behavior that can affect the model, one important result of this paper is that the analytic signal derived from the decomposed SoH provides means to detect changes i.e., it provides means to analyze if the SoH starts to deteriorate or to change less frequently or more frequently than before. This information may indicate a need to revamp the model. This conclusion required discussions with mechanical engineers who pointed out that the deterioration of a lithium-ion battery tends to accelerate in the very end-of-lifetime, something which our model was not able to learn with only the few first years of data collected from batteries for model development. There was no significant change in the deterioration rate during the observation period, although some interesting temporary changes were detected, probably due to turning off sensory equipment of the battery. It is for future work to implement an integration of this instantaneous frequency with automatic analysis functionality into the machine learning model.

To make a baseline model, we utilized ARIMA method to create one. Having the assumption of downward trend in mind, we derived a model ARIMA(0,1,1) that yielded the RMSE for SoH of 2.95. The ARIMA(0,1,1) was an exponential smoothing; and the results reasonable; however, the

RMSE is relatively big as it averages the results considerably. Therefore, it is fit for the purpose of overall detection of a SoH degradation trend, but not to predict minute SoH values. See Figure 3.4.



**Figure 3.4.** All SoH values together with ARIMA (0,1,1) predicted SoH values.

## 4. Article III: event detection in imbalanced data

This chapter sets to resolve application challenges introduced in Chapter 1 that are related to imbalanced data and the imperfect data recordings. The imbalanced and imperfect data challenges are a part of the second research question (RQ2) identified foremost in Publication III.

### 4.1 Background

More data is available from buildings than ever. This is a prerequisite to creating smart building energy applications; however, at the same time, a drawback with lots of this data is that it is often imbalanced, incomplete, and even unlabeled sometimes. An example of imbalanced data is classification data with skewed class proportions. That is, an imbalanced classification data set consists of classes with the majority proportion of data (called the majority classes) and the minority proportion (called the minority classes). Changing the composition of a training data set for an imbalanced classification task addresses the problem of imbalanced data. For this aim, resampling methods are utilized to change the composition where pieces of data are deleted from the transformed data, pieces of data are copied and supplemented to the transformed data, or both of these two methods are combined. Undersampling refers to a group of heuristic techniques balancing the class distribution by deleting majority class examples from data set that has a skewed class distribution in a defined manner. In turn, an example of incomplete data is a time series that lacks periods of data in the data set. This may or may not change the class distribution; therefore, it may affect the results of resampling and classification. This case-study proposes a novel data undersampling algorithm to be utilized together with other resampling algorithm to preprocess the data. The algorithm was verified with a classification task in the presence of imbalanced and missing data.

## 4.2 Case-study methods

In our case-study, both TomekLinks undersampling and an undersampling algorithm proposed are used in a combination with a Logistic Regression to make binary classification of the data.

TomekLinks undersampling detects relationship between samples [44]. A relationship, so called Tomek’s link, between a sample  $x$  of a class and a sample  $y$  of another class is defined such that for any sample  $z$ :

$$d(x, y) < d(x, z) \text{ and } d(x, y) < d(y, z) \quad (4.1)$$

where  $d(\cdot)$  is the distance between the two samples so that there is a Tomek’s link between the two samples if they are the nearest neighbors of each other. After identification of the link, a sample belonging to the majority class can be remove, that is, the majority class is undersampled to have more balance ratio between the majority and minority classes. The undersampled data is then utilized by a classification method, such as Logistic Regression that is presented in brief below.

For an individual observation  $y_i \in 0, 1$  related to covariates  $x_i \in \mathbb{R}^p$ , we assume that  $Y_i \sim \text{Bernoulli} \in 0, 1$ . The parameter  $\pi_i$  is derived as follows:

$$\pi_i = \sigma(x_i) = \frac{1}{1 + \exp(-\phi(x_i))}, \quad (4.2)$$

where  $\phi(x_i)$  is the predictor function  $\mathbf{w}^T X - i$ ;  $\mathbf{w} \in \mathbb{R}^p$ . The values of the parameter  $\mathbf{w}$  can be estimated by minimizing the negative log likelihood of the chosen Bernoulli data model.

$$\text{loss} = - \sum_{i=1}^n y_i [\log(\sigma(x_i)) + (1 - y_i) \log(1 - \sigma(x_i))], \quad (4.3)$$

To classify events reliably, we formulated a hypothesis that the classification result may be improved by pre-processing the data utilizing the temporal characteristics of the observations. To this end, we now present an algorithm (Algorithm 1) to undersample the majority class data in proximity of the minority class data as an extension to the strategy proposed in [45]. To ensure the randomness to to avoid overfitting, the algorithm is applied only to the data used for learning the classification model; moreover, the data is repeatedly and randomly split before it is used by the learning algorithm. More specifically, all data is split randomly  $n_i$ -fold so that the resulting train and validation sets contain 50% of the one-year worth of data; this is used as the means for stratifying true events to both train and validation sets, albeit in crude yearly granularity. For each fold, the number of  $n_u$  events preceding a true event is scrutinized and false events, that are the majority class in this case, are deleted from these  $n_u$  events. That is, the algorithm undersamples the majority class samples prior to a minority class sample in the time series manner so that the maximum

number of  $n_u$  prior false events are deleted. For each fold,  $n_i$ , the algorithm returns the undersampled validation set  $UTr$  for cross-validation purposes (Algorithm 2). In the model proposed, to the Logistic Regression classification. The finally developed model was verified with a test set (unseen data for model).

---

**Algorithm 2** Undersample timewindows with Boolean masks for cross-validation purposes

---

**Input:** ( $Awin, n_u$ )  
**Output:** Undersampled( $UTr$ ) and GroundTruth( $Te$ )  
**1:**  $Tr, Te \leftarrow \text{Split}(Awin)$   
**2:**  $m_1, \dots, m_{n_u} \leftarrow [], \dots, []; i_m \in \{True, False\}$   
**3:**  $m_1 \leftarrow \text{GetTarget}(Tr[1], \dots, Tr[|Tr|])$   
**4:**  $m_2, \dots, m_{n_u} \leftarrow False$   
**5:** **for**  $U \leftarrow n_u + 1$  **to**  $|Tr - n_u|$  **do**  
**6:**     **if**  $m_1[U] == True$  **then**  
**7:**          $m_2[U - 1], \dots, m_{n_u}[U - n_u] \leftarrow True$   
**8:**     **end if**  
**9:** **end for**  
**10:**  $m_1 \leftarrow (m_1)^C$   
**11:**  $m_2, \dots, m_{n_u} \leftarrow m_1$  **and**  $m_2, \dots, m_1$  **and**  $m_{n_u}$   
**12:**  $UTr \leftarrow Tr \setminus Tr[m_2 \text{ or } \dots \text{ or } m_{n_u}]$   
**13:** **return**  $UTr, Te$   
**14:**  $\rightarrow \text{CrossValidate}(UTr, Te)$

---

To define a metrics for compare the the model results, we use a F0.5-measure. Generically, the cost associated with an inaccurate prediction depends on the number of false positives according to the cost function for optimal prediction:

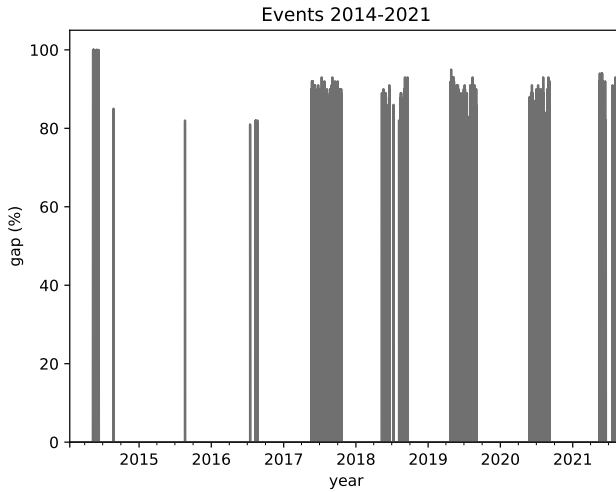
$$\sum_j P(j|x) Cost(i, j, x) \quad (4.4)$$

where prediction  $x$  is associated with function  $Cost(i, j, x)$ , which defines the cost of predicting class  $i$  for  $x$  when the true class is  $j$ . The goal of equation (4.4) is to determine how to minimize the expected cost; the cost function that we utilized was the F0.5-measure that puts more weight on precision than recall. Precision and recall are expressed in scale of 0-1, 1 being the best score.

### 4.3 Case-study implementation

To evaluate the model comprising of TomekLinks and Algorithm 2 and Logistic Regression, we executed a basic comparison of several oversampling, undersampling and machine learning models to find the best model in terms of F0.5-metrics.

The events that we were trying to classify were imbalanced (5.3% of data were labeled as positive events), and distributed unevenly in the time-series (Figure 4.1).



**Figure 4.1.** Recorded events in 2014—2021.

#### 4.4 Discussion of undersampling algorithm and event detection case-study

To verify the best selected model Logistic Regression with TomekLinks and Algorithm 2, this final model was verified with the unseen data yielding F0.5 measure of 0.648 (0.019), the precision of 0.93 and the recall of 0.55. The selected metrics emphasizes the ratio of true positives vs. false positives. Therefore, the final model fulfilled the set goal: we aimed at a high precision, while a low accuracy was acceptable. These results are satisfactory taking into account the three major factors in the model development as follows: the data was imbalanced (containing  $< 5.3\%$  target events), the data distribution was imbalanced both within a year (approximately March-October in Fig. 4.1) and year-on-year. Lastly, the equipment was turned off during periods that overlapped with the target events; moreover, the missing data contains  $8.3\%$  of data. This reflects the importance of systematic collection of data, and labeling the data.

The data collected warrants optimization of heat recovery functionality as the faults or imperfect heat recovery periods, lasted  $6\%$  of the seven-year data, sometimes lasting for several days. Sometimes this caused discomfort to the building users that we found out in the related discussions. The proposed model gives us a tool to initiate more analysis with the help of classifying events, which enables labeling of observations with more detailed classification, such as in [46]. Consequently, our model can be a building block for a new model with more precision and accuracy to forecast several types of events, for instance, to correct errors.

## 5. Article IV: thermal preference classification

This chapter sets to resolve application challenges introduced in Chapter 1 that are related to imbalanced data and the human-in-the loop for smart building applications. The direct involvement of people in the thermal adjustments for a room is a challenge that is a part of the third research question (RQ3) identified foremost in Publication IV.

### 5.1 Background

Thermal comfort is a key aspect for enhancing the indoor environment quality and the energy efficiency of buildings. In modern office buildings today, the temperature of office rooms is typically set to a value and kept there with building automation even though number of people changes during a day. For example, the value 21.5 °C is widely utilized in the Nordic countries. However, building automation can also control different parts of a building in different manner if it is built in flexible manner, for example, with individual coolers in each office room.

There are two main methods for advanced building controls to assess thermal comfort indoors: the thermal balance method and the adaptive method. The object of the thermal balance method is to keep body temperature in balance in the indoor environment; this method takes into account factors such as air temperature, relative humidity, air velocity and air radiant temperature, metabolic activity and clothing. This method, overall, represents the human as a passive actor who absorbs the effects of an environment. In contrast to this, the adaptive method represents, overall, the human as an adaptive actor. In this latter method, factors such as human age, gender, pathologies, current activity and outdoor temperature can be taken into account.

The thermal comfort is typically indicated with a number derived from user-studies. This number presents a mean of the opinions with given environmental factors. However, applying this thermal assessment into building controls remains a challenge. Furthermore, personalizing the

assessment model is another notable challenge. The aim of this paper is to yield few multi-class estimate of the personal thermal preferences utilizing sensors in the building, on person, and also ask about the comfort directly with a simple question. The proposed model extends the present thermal comfort assessment methods by yielding multi-class estimations. This multi-class estimate could be utilized to enhance the advanced building controls to better match the temperature values in a building to the preferences of its present occupants.

## 5.2 Related work

A literature review on the subject indicated that recent papers model the thermal comfort with machine learning methods based on the balance method [47, 48]. In contrast to the material reviewed, this paper expands the findings of the previous research on adaptive methods in both comparing more methods and in relying more on the time-dependency of the data [49, 50]. Finally, we wanted to develop a model that generalizes reasonably to a set of people in a given environment, contrary to developing individual models as, e.g., is the case in paper [51].

## 5.3 Case-study methods

In order to retrieve the sensor data from this AHU, we employed a local open messaging interface (O-MI) node to collect the data and utilized the open data format (O-DF) protocol to transfer the data over the network as the building equipment is not connected to the open internet [14, 15].

To make multi-class classification, we selected one-versus-all strategy for multi-class classification with random forest as the base classifier [35]. In one-versus-all strategy a sample was classified to belong to one of the three classes defined (prefers cooler, no change, prefers warmer). Then, according to the selected one-versus-all strategy, multiple binary classification models were fitted for each class vs. all other classes to find the best model utilizing the base classifier.

## 5.4 Case-study implementation

To develop a personalized morel, the aim of this paper is to yield an estimate of personal thermal preference utilizing sensors in the building, on person and also to ask about the comfort directly with a simple question. This assessment can be utilized with advanced building controls including this kind of model.

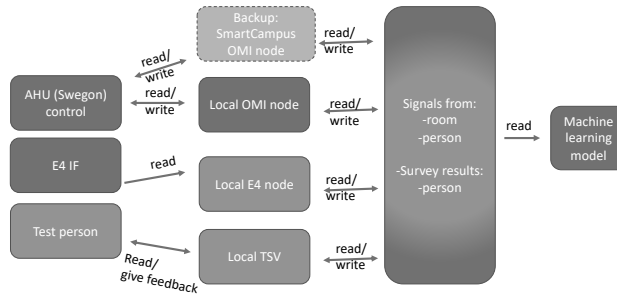
Ten subjects (4 females, 6 males) living in the metropolitan Helsinki area, Finland, were recruited through personal contacts. All subjects were working or studying in the Helsinki area. Table 5.1 describes the data on the subjects during the experiment period. The experiment was repeated six times in different temperatures and on different activity levels (resting, exercise) as a proxy for entering the room for work. Due to Covid-19 restrictions we were unable to expand the number of people involved.

**Table 5.1.** Summary of the experiment subjects for evaluation of personal thermal comfort in an office room.

ID	sex	age decile
1	M	5
2	M	2
3	M	3
4	W	3
5	M	1
6	W	5
7	M	5
9	M	5
10	W	3
11	W	2

To collect physiological data on subjects, we utilized a wearable wristlet device (Empatica E4).

The experiment was conducted at Aalto University premises (Espoo, Finland) in an office room. To measure the temperature, humidity, and CO<sub>2</sub> levels of the room, we utilized the sensors of the integrated air handling unit (AHU, manufactured by Swegon AB), which was located 2 m above the table where experiment was conducted (Figure 1).



**Figure 5.1.** The overall architecture of the comfort model case-study.

## 5.5 Case-study discussion

The models one-versus-the-rest combined with Random Forests yielded promising results to estimate the personal thermal preference classes.

The models generalized well to the test data as the yielded F1-measure (micro) 0.80–0.83 indicates that the models perform satisfactorily on new data (new subjects). The data for these models was collected at specific points of time (at around 20 minutes and after standardized exercise). We noted that it is debatable how well the experiment duration, the varied temperature range and the defined exercise regime during experiment managed to capture thermal preference during a working day in an office. One of the conclusions was that the result of model classification depends largely on the quality and quantity of features, not the performance of the model itself.

Despite the caveats mentioned, we feel that the yielded thermal preference model gives a tool to extract personal thermal comfort for building controls. Furthermore, it can be easily turned into recommendations for the building users, and into aids to adjust the building temperature settings of different rooms, for instance.

## 6. Article V: applications for smart building energy utilization

This chapter focuses on resolving important application challenges introduced in Chapter 1: what are the elements needed for the smart building energy applications? This is a part of the fourth research question (RQ4) examined in Publication V. Furthermore, this also combines the works in Publications I-IV to create context for the smart building energy utilization.

Recently, the industry has drawn attention to the Smart Grid applications, but deployment of many other application areas have been laggard. Moreover, the relationship between the different domains are sometimes overlapping but sometimes far apart. A systematic approach to classify these applications also helps to understand these relationships. This helps the integration of disparate models that is essential to fully exploit the models in buildings.

### 6.1 Background

Overall, households and transport account for over 60% of energy consumption in Europe and North America; this share is 40% and rising in Asia [52]. Furthermore, 40–50% of this energy is directed to HVAC in Northern Europe and the USA [2, 53]. To promote efficient energy consumption, there is a need to develop and apply machine learning (ML) applications coupled with systematic data.

The absence of applicable data for buildings and people is perhaps the primary problem when attempting to create a solution. However, to form a common ground and to improve the research preconditions, some preliminary efforts exist to provide more open data. These include, for instance, the ASHRAE Global Thermal Comfort Database [54] on personal behavior and comfort data, and the Building Data Genome Project [55] on building data.

To address these challenges, a primary solution is to deploy applications and, simultaneously, ensure that the heterogeneous data is adequately

acquired and processed to fit the needs of applications. Many smart energy applications for buildings necessitate a data collection and processing in a way required by the methods. The examples include imputing missing observations to complete a data series, or the data transformations required to achieve the results desired from the algorithms utilized. At times, more sensors and ways to interact with the people and environment are needed. To these ends, the solutions for buildings should provide the following characteristics: (i) a systematic approach to identify the need and related requirements for an application (ii) the fulfillment of these often parallel needs for building energy management, smart grid, personalization, and maintenance and security; and (iii) the capability to processing data adequately.

## 6.2 Related work

According to the most cited journal papers during the years 2009-2021, energy-efficiency, user-comfort, and maintenance and security applications comprise the smart building energy applications. An overview of the related surveys on the topic is in Table 6.1, which is arranged according to a reference number, publications year, requirement class, short description, main challenges, and methodology for the utilized taxonomy. According to the short description column, the surveys constitute various application areas. Hence, also the problems faced are heterogeneous, albeit in the taxonomy proposed.

The table is arranged according to the survey citation number, the publication year, key search phrases, the application domain, issues and challenges, and the classification proposed.

**Table 6.1.** Identified key surveys for building applications.

No.	Yr.	Key phrase(s)	Application domain	Issues, challenges	Classification
[56]	2009	efficiency; comfort	building energy management	11 issues related to building and ML	classification to 17 control methods
[57]	2014	efficiency; comfort	building energy management	energy unaware activities	5-class control system classification
[3]	2015	efficiency; comfort	building energy management	six issues of model development	histogram of home appliances vs. surveys
[58]	2015	efficiency	microgrids	eight issues of microgrids	8-class issue based classification
[59]	2015	comfort	building energy management	scheduling charging of electric vehicles	flowchart of EV scheduling objectives and methods
[60]	2015	efficiency	microgrids	energy storage techniques vs. costs	two flowcharts for energy storages; several cost histograms
[61]	2016	efficiency	microgrids	energy sharing, management	classification
[62]	2016	efficiency	smart grid	12 technical and economical issues	comparison of grid types; classification of grid domains (users, electricity generation, etc.)
[63]	2017	efficiency; comfort	building IoT	7-leaf daisy chart on issues	4-layer taxonomy (applications and frameworks)
[64]	2017	efficiency	building energy management	data collection from buildings	classification to nine forecast methods
[65]	2017	efficiency	microgrids	more implementations needed	categorizing renewable energy with utilized ML methods

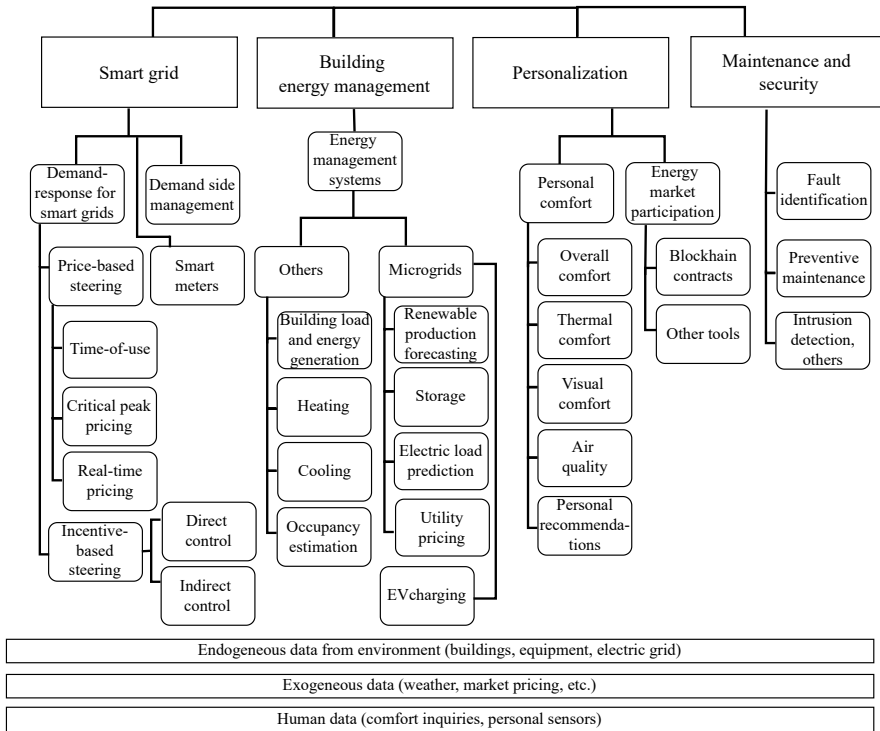
[66]	2017	efficiency	smart grid	heterogeneous buildings; varying occupancy characteristics	flowcharts of smart meter models
[67]	2018	efficiency; comfort	building energy management	models limited to training data; understanding black-box models	five pie-charts on model (algorithms etc.)
[68]	2018	efficiency; comfort	smart grids	volatility of renewable energy; occupant behavior	classification to demand and supply side
[69]	2018	anomaly	maintenance	ML model generalization; limited reproducibility	2-layer daisy-chart categorization of applications
[70]	2018	efficiency; comfort	personalization	triggering relevant control action	3-class classification of building controls
[71]	2018	efficiency	smart grid	multiple domains; conflicting objectives	flow-charts for energy modeling
[72]	2018	efficiency	microgrids	microgrid reliability, security, costs	9-class microgrid classification
[73]	2019	efficiency	data	issues of system, data management, and analysis	2 classifications on data
[74]	2019	anomaly	security	proper operation of grid and data security	data security, classification of applications and data, mapping to ML methods
[75]	2019	efficiency; comfort	personalization	energy-efficiency and user comfort balance	5-layer hierarchical taxonomy
[76]	2019	efficiency; comfort	data	data storing, interpreting black-box models	flowchart for a energy control system
[77]	2019	anomaly	security	seven challenges; hashing of blockchains may be unraveled by quantum computers	3-layer application taxonomy
[78]	2019	anomaly; efficiency	smart grid, security	several ML and system issues	4-layer smart meter taxonomy
[79]	2020	efficiency; anomaly	smart grid; building energy	complex sequential decision-making	frameworks for systems, flowchart for algorithms
[80]	2020	anomaly	security	power system security and stability	3-layer security assessment, 4-layer classification
[81]	2020	anomaly; efficiency	smart grid	integration of building systems and security	framework and classification
[82]	2020	efficiency	smart grid	incentives, usage patterns, scheduling	2-category incentive classification of energy markets
[83]	2020	anomaly; efficiency	data, building load	data including both aggregated and appliance level for models	3-layer taxonomy
[84]	2020	comfort	maintenance	limited model transferability, data, costs	6-phase lifecycle
[85]	2020	anomaly	security, data	six AI methods and model challenges	several AI technique classifications
[12]	2021	efficiency; anomaly	building energy management	selecting database, consumption profiles of buildings	4-layer daisy chart taxonomy
[86]	2021	anomaly; comfort	maintenance	changes in building configuration	4 energy management strategies, classifications
[87]	2021	efficiency	personalization	throughput, consensus contract mechanisms, security	10 blockchain objectives, classifications
[88]	2021	anomaly	smart grid; security	data formats, privacy	3-layer service classification
[89]	2021	efficiency	security; building energy	data quality, security, lack of ML tools	3-classes of machine learning modeling

According to the application domain column, the surveys constitute various application areas. We retrieved surveys related to smart grid, several types of building energy control and management systems, personalization, and maintenance and security. Moreover, the issues and challenges column indicated that problems faced are heterogeneous. In the following, the application domain areas are discussed in more detail below before a novel taxonomy is proposed.

### 6.3 Taxonomy for smart building energy utilization

As a summary, these and other solutions in the reviewed papers utilize several approaches to formulating the taxonomy, for instance, based on a framework for workflow. Some surveys concentrate on a type of machine learning, such as, neural networks [79], or on an application area, for

instance energy efficiency, such as [89]. Moreover, [84] presents operation, control and retrofit; however, it does not include the human-in-the-loop as such in its taxonomy. In contrast, for example, [90] presents solutions that incorporate occupants of the buildings in sensory and control frameworks. To encounter these challenges, we proposed a classification for the smart building energy utilization in Paper V. The classification has four main categories: building energy management, smart grid, personalization, and maintenance and security. For the complete taxonomy, see Fig. 6.1.



**Figure 6.1.** Taxonomy of applications for smart building energy utilization.

The solution proposed for state-of-health for lithium-ion batteries in Publications I-II belong to the category of maintenance. The imbalanced data processing in Publication III belongs primarily to all of the application categories through the ubiquitous basis category of endogenous data from the environment. The personal thermal comfort application in Publication IV belongs to category of thermal comfort. It is also an example of an application belonging to the maintenance category.

The taxonomy proposed helps to understand the relationships between the applications. This is a tool to plan and integrate of disparate models that is essential to exploit the models fully in buildings. It also presents how many different components are needed to build an application, or in general, an environment for the applications.

## 7. Conclusions

The basis for present day energy systems in most countries is simple scheme, where energy resources are converged to meet the demand. Moreover, increased demand is met with increased utilization of resources [91]. Recent applications provide sustainable energy solutions for the energy utilization of buildings to a growing extent. These applications enable the utilization of renewable energy sources, energy storages and national or local energy systems in an energy-efficient and flexible manner in contrast to the traditional systems. They also satisfy the needs for comfort, maintenance and security that form an integral part of the solutions. This dissertation addresses the issue of smart building energy utilization from two major aspects: creation of a taxonomy and selected case-studies of applications. In this context, several solutions, and two algorithmic approaches are proposed based on the heterogeneous data available.

First, we examine how to estimate the aging of the equipment and the uncertainty related the estimations for an implementation of smart building energy application. In the Publications I and II we investigated how to ensure safety of some smart building applications (RQ1). In the case of battery equipment this is imperative as the battery may self-ignite as it wears out. We looked at one battery and generalized the results successfully to several batteries. Furthermore, we proposed a walk-forward algorithm that can flexibly use the data one time-step at a time, or in batches of several time-steps to forecast and evaluate the forecast. Not limited to lithium-ion battery state-of-life, this method is applicable to other same kind of regression problems.

Second, we investigate how to build a model in case of imbalanced data with missing observations (RQ2). To enhance undersampling of this kind of data, we proposed a novel data undersampling algorithm. We utilized this data with classification models to compare and verify the results. The combined model utilizing undersampling algorithms and machine learning classification developed in Publication III yielded satisfactory results despite the data imbalance, uneven distribution and missing data. We successfully met the goal of creating model that predicted events with

a high precision.

Third, we address the need to involve people in environmental comfort decisions directly (RQ3). To personalize the thermal adjustments in buildings, we proposed a model to estimate personal thermal preference utilizing sensors in the building as well as on person and also ask about the comfort directly with a simple question in Publication IV. We were able to create a model that satisfactorily forecast a personal thermal comfort class. The results can be easily turned into recommendations for the building users to be used to select a room to work in, or to adjust building temperature settings.

Finally, in Publication V we survey the application in the field of smart building energy utilization to form an overview and create a taxonomy of these applications (RQ4). The relationship between the different domains are sometimes overlapping but sometimes far apart. A systematic approach to classify these applications also helps to understand these relationships. This promotes the integration of disparate models that is essential to utilize the models fully in buildings.

In a summary, we resolved issues related to overall classification of the applications. In addition, we proposed solutions and algorithm that can be applied to model development even outside the scope of this case-study.

## 7.1 Implications and limitations

In the beginning of this dissertation, we specified some major challenges recently faced by the application development for smart building energy utilization. We proposed several applications and solutions in the dissertation. In addition, the solutions were validated through real-life case studies that implied that the author managed to contribute to build the targeted applications. In some cases, the research appears to improve the existing methodologies with enhancements to algorithms with generic applicability.

*The state-of-health* is essential for the safety of the energy storage applications. We proposed two different models (ARIMA and Gradient Boosting based models). The limitation is related to the mechanical aspects of the equipment used, as the model is heavily dependent on the data particular to this type of batteries.

*Imbalanced data* requires mitigating efforts to employ this data in model creation and forecasting. However, the data suffers from missing labels, so the algorithm proposed is on very generic level. It addresses the basic problem of imbalanced data, but does not resolve the actual problem of data quality for detecting faults or other items of interest accurately.

*Personal comfort* is a subjective matter. The major limitation is related to the need to collect not only the human signals minutely (e.g., with a

wristlet attached) but also some personal demographic information, like gender and age, to make an accurate forecast. Although the big data companies can collect enough data, this will be a struggle for companies in the field of energy and buildings.

Any *taxonomy* has to be precise to be useful; however, there is an obvious tendency in the present research to overlap with the aspects of smart energy and personalization, which makes it difficult to, for instance, compare the different taxonomies.

## 7.2 Future directions

The applications discussed in this dissertation are part of the latest developments for the smart buildings. How to actually integrate these, or even other, applications, especially in the field of personalization, remains an open issue.

Although this dissertation addressed several of the issues related to present-day solutions, there are many apparent problems that remain. The absence of applicable data is perhaps the primary issue. This includes both the absence of public databases to make models, and the limited availability of appropriate sensor data from buildings. The absence of well-defined personal data is another issue when making applications due to perceived difficulty of measuring user preferences and behavior. Integration of the applications into coherent solutions is another problem that needs to be addressed to increase the utilization and benefits from the applications. These problems give a direction where to focus in future research.



# References

- [1] Henrik Lund, Poul Alberg Østergaard, David Connolly, and Brian Vad Mathiesen. Smart energy and smart energy systems. *Energy*, 137:556–565, 2017.
- [2] Jose Lisandro Aguilar Castro, Alberto Garcés Jiménez, María Dolores Rodríguez Moreno, Rodrigo García, et al. A systematic literature review on the use of artificial intelligence in energy self-management in smart buildings. *Renewable and Sustainable Energy Reviews*, 151:111530, 2021.
- [3] Marc Beaudin and Hamidreza Zareipour. Home energy management systems: A review of modelling and complexity. *Energy Solutions to Combat Global Warming*, pages 753–793, 2017.
- [4] United Nations Department of Global Communications. The 17 goals. Online, 2019. 2015-09-18.
- [5] The European Parliament and the Council. Directive (EU) com/2021/557, 2021.  
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0557>.
- [6] Asad Javed, Sylvain Kubler, Avleen Malhi, Antti Nurminen, Jeremy Robert, and Kary Främling. Biotope: building an iot open innovation ecosystem for smart cities. *IEEE Access*, 8:224318–224342, 2020.
- [7] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.
- [8] Kai Ming Ting. Precision and recall. *Encyclopedia of machine learning*, 781, 2010.
- [9] Thomas J Santner, Brian J Williams, William I Notz, and Brian J Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003.
- [10] Jason Runge and Radu Zmeureanu. Forecasting energy use in buildings using artificial neural networks: A review. *Energies*, 12(17):3254, 2019.
- [11] Mengjie Han, Ross May, Xingxing Zhang, Xinru Wang, Song Pan, Da Yan, Yuan Jin, and Liguu Xu. A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustainable Cities and Society*, 51:101748, 2019.
- [12] Yassine Himeur, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Applied Energy*, 287:116601, 2021.

- [13] Roy Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvin, 2000. <https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>.
- [14] The Open Group. Open messaging interface (o-mi). Online, 2019.
- [15] The Open Group. Open data format (o-df). Online, 2019.
- [16] Kary Främling, Sylvain Kubler, and Andrea Buda. Universal messaging standards for the iot from a lifecycle management perspective. *IEEE Internet of things journal*, 1(4):319–327, 2014.
- [17] Sylvain Kubler, Jérémy Robert, Ahmed Hefnawy, Kary Främling, Chantal Cherifi, and Abdelaziz Bouras. Open iot ecosystem for sporting event management. *IEEE Access*, 5:7064–7079, 2017.
- [18] Kary Främling, Jan Holmström, Timo Ala-Risku, and Mikko Kärkkäinen. Product agents for handling information about physical objects. *Report of Laboratory of information processing science series B, TKO-B*, 153(03), 2003.
- [19] Kary Främling, Jan Holmström, Juha Loukkola, Jan Nyman, and André Kaustell. Sustainable plm through intelligent products. *Engineering Applications of Artificial Intelligence*, 26(2):789–799, 2013.
- [20] Copenhagen Centre on Energy. Fit for 55. Online, 2022.
- [21] Kaile Zhou and Shanlin Yang. Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renewable and Sustainable Energy Reviews*, 56:810–819, 2016.
- [22] Christian F Calvillo, Alvaro Sánchez-Miralles, and Jose Villar. Energy management and planning in smart cities. *Renewable and Sustainable Energy Reviews*, 55:273–287, 2016.
- [23] Egoitz Martinez-Laserna, Elixabet Sarasketa-Zabala, D-I Stroe, M Swierczynski, Alexander Warnecke, Jean-Marc Timmermans, Shovon Goutam, and Pedro Rodriguez. Evaluation of lithium-ion battery second life performance and degradation. In *2016 IEEE Energy Conversion Congress and Exposition (ECCE)*, pages 1–7. IEEE, 2016.
- [24] James D Kozlowski. Electrochemical cell prognostics using online impedance measurements and model-based data fusion techniques. In *2003 IEEE Aerospace Conference Proceedings (Cat. No. 03TH8652)*, volume 7, pages 3257–3270. IEEE, 2003.
- [25] Yin Hua, Andrea Cordoba-Arenas, Nicholas Warner, and Giorgio Rizzoni. A multi time-scale state-of-charge and state-of-health estimation framework using nonlinear predictive filter for lithium-ion battery pack with passive balance control. *Journal of Power Sources*, 280:293–312, 2015.
- [26] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
- [27] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [28] Jörn Vogel, Claudio Castellini, and Patrick van der Smagt. Emg-based teleoperation and manipulation with the dlr lwr-iii. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 672–678. IEEE, 2011.

- [29] Yuhang Ye, Chao Liu, Nabil Zemiti, and Chenguang Yang. Optimal feature selection for emg-based finger force estimation using lightgbm model. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–7. IEEE, 2019.
- [30] Mohamed M Ahmed and Mohamed Abdel-Aty. Application of stochastic gradient boosting technique to enhance reliability of real-time risk assessment: use of automatic vehicle identification and remote traffic microwave sensor data. *Transportation research record*, 2386(1):26–34, 2013.
- [31] Yung-Chia Chang, Kuei-Hu Chang, and Guan-Jhih Wu. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, 73:914–920, 2018.
- [32] Fu-Kwun Wang and Tadele Mamo. Gradient boosted regression model for the degradation analysis of prismatic cells. *Computers & Industrial Engineering*, 144:106494, 2020.
- [33] Miloš Polák and Lenka Drápalová. Estimation of end of life mobile phones generation: The case study of the czech republic. *Waste management*, 32(8):1583–1591, 2012.
- [34] Claude Sammut and Geoffrey I. Webb, editors. *Mean Absolute Error*, pages 652–652. Springer US, Boston, MA, 2010.
- [35] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning The Elements of Statistical Learning, vol. 27*. Springer series in statistics New York, 2017.
- [36] MATLAB. Rolling-window analysis of time-series models, 2021.
- [37] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971):903–995, 1998.
- [38] N Ramesh Babu and B Jagan Mohan. Fault classification in power systems using emd and svm. *Ain Shams Engineering Journal*, 8(2):103–111, 2017.
- [39] Claude Sammut and Geoffrey I. Webb, editors. *Mean Squared Error*, pages 653–653. Springer US, Boston, MA, 2010.
- [40] George Hebrail and Alice Berard. Individual household electric power consumption data set, 2010.
- [41] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system acm sigkdd international conference on knowledge discovery and data mining. *ACM*, pages 785–794, 2016.
- [42] Denise Rey and Markus Neuhäuser. *Wilcoxon-Signed-Rank Test*, pages 1658–1659. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
- [43] Uber Engineering. Building a backtesting service to measure model performance at uber-scale, 2021.
- [44] Ivan Tomek. Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772, 1976.
- [45] Luis Torgo, Paula Branco, Rita P Ribeiro, and Bernhard Pfahringer. Resampling strategies for regression. *Expert Systems*, 32(3):465–476, 2015.

- [46] Manik Madhikermi, Avleen Kaur Malhi, and Kary Främling. Explainable artificial intelligence based heat recycler fault detection in air handling unit. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 110–125. Springer, 2019.
- [47] Ghezlane Halhoul Merabet, Mohamed Essaaidi, Mohamed Ben Haddou, Basheer Qolomany, Junaid Qadir, Muhammad Anan, Ala Al-Fuqaha, Mohamed Riduan Abid, and Driss Benhaddou. Intelligent building control systems for thermal comfort and energy-efficiency: A systematic review of artificial intelligence-assisted techniques. *Renewable and Sustainable Energy Reviews*, 144:110969, 2021.
- [48] Adam Rysanek, Rohan Nuttall, and Justin McCarty. Forecasting the impact of climate change on thermal comfort using a weighted ensemble of supervised learning models. *Building and Environment*, 190:107522, 2021.
- [49] Da Li, Carol C. Menassa, and Vineet R. Kamat. Personalized human comfort in indoor building environments under diverse conditioning modes. *Building and Environment*, 126:304–317, 2017.
- [50] Francesco Salamone, Lorenzo Belussi, Cristian Currò, Ludovico Danza, Matteo Ghellere, Giulia Guazzi, Bruno Lenzi, Valentino Megale, and Italo Meroni. Integrated method for personal thermal comfort assessment and optimization through users’ feedback, iot and machine learning: A case study. *Sensors*, 18(5):1602, 2018.
- [51] Shichao Liu, Stefano Schiavon, Hari Prasanna Das, Ming Jin, and Costas J. Spanos. Personal thermal comfort models with wearable sensors. *Building and Environment*, 162:106281, 2019.
- [52] anon. *2021 Energy Statistics Pocketbook*. Series E No.4. United Nations, New York, 2021.
- [53] Joon-Ho Choi Choi and Dongwoo Yeom. Study of data-driven thermal sensation prediction model as a function of local body skin temperatures in a built environment. *Building and Environment*, 121:130–147, 2017.
- [54] Veronika Födová Ličina, Toby Cheung, Hui Zhang, Richard de Dear, Thomas Parkinson, Edward Arens, Chungyoon Chun, Stefano Schiavon, Maohui Luo, Gail Brager, et al. Development of the ashrae global thermal comfort database ii. *Building and Environment*, 142:502–512, 2018.
- [55] Clayton Miller and Forrest Meggers. The building data genome project: An open, public data set from non-residential building electrical meters. *Energy Procedia*, 122:439–444, 2017.
- [56] Anastasios I Dounis and Christos Caraiscos. Advanced control systems engineering for energy and comfort management in a building environment—a review. *Renewable and Sustainable Energy Reviews*, 13(6-7):1246–1261, 2009.
- [57] Pervez Hameed Shaikh, Nursyarizal Bin Mohd Nor, Perumal Nallagownden, Irraivan Elamvazuthi, and Taib Ibrahim. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renewable and Sustainable Energy Reviews*, 34:409–429, 2014.
- [58] Sina Parhizi, Hossein Lotfi, Amin Khodaei, and Shay Bahramirad. State of the art in research on microgrids: A review. *Ieee Access*, 3:890–925, 2015.
- [59] Zhile Yang, Kang Li, and Aoife Foley. Computational scheduling methods for integrating plug-in electric vehicles with power systems: A review. *Renewable and Sustainable Energy Reviews*, 51:396–416, 2015.

- [60] Behnam Zakeri and Sanna Syri. Electrical energy storage systems: A comparative life cycle cost analysis. *Renewable and sustainable energy reviews*, 42:569–596, 2015.
- [61] Aftab Ahmad Khan, Muhammad Naeem, Muhammad Iqbal, Saad Qaisar, and Alagan Anpalagan. A compendium of optimization objectives, constraints, tools and algorithms for energy management in microgrids. *Renewable and Sustainable Energy Reviews*, 58:1664–1683, 2016.
- [62] Maria Lorena Tuballa and Michael Lochinvar Abundo. A review of the development of smart grid technologies. *Renewable and Sustainable Energy Reviews*, 59:710–725, 2016.
- [63] Mussab Alaa, Aws Alaa Zaidan, Bilal Bahaa Zaidan, Mohammed Talal, and Miss Laiha Mat Kiah. A review of smart home applications based on internet of things. *Journal of Network and Computer Applications*, 97:48–65, 2017.
- [64] Chirag Deb, Fan Zhang, Junjing Yang, Siew Eang Lee, and Kwok Wei Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.
- [65] Sunil Kr Jha, Jasmin Bilalovic, Anju Jha, Nilesh Patel, and Han Zhang. Renewable energy: Present research and future scope of artificial intelligence. *Renewable and Sustainable Energy Reviews*, 77:297–317, 2017.
- [66] Baran Yildiz, Jose I Bilbao, Jonathon Dore, and Alistair B Sproul. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy*, 208:402–427, 2017.
- [67] Kadir Amasyali and Nora M El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2018.
- [68] Yongbao Chen, Peng Xu, Jiefan Gu, Ferdinand Schmidt, and Weilin Li. Measures to improve energy demand flexibility in buildings for demand response (dr): A review. *Energy and Buildings*, 177:125–139, 2018.
- [69] Clayton Miller, Zoltán Nagy, and Arno Schlueter. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews*, 81:1365–1377, 2018.
- [70] Sophie Naylor, Mark Gillott, and Tom Lau. A review of occupant-centric building control strategies to reduce building energy use. *Renewable and Sustainable Energy Reviews*, 96:1–10, 2018.
- [71] Lamberto Tronchin, Massimiliano Manfren, and Benedetto Nastasi. Energy efficiency, demand side management and energy storage technologies—a critical analysis of possible paths of integration in the built environment. *Renewable and Sustainable Energy Reviews*, 95:341–353, 2018.
- [72] Muhammad Fahad Zia, Elhoussin Elbouchikhi, and Mohamed Benbouzid. Microgrids energy management systems: A critical review on methods, solutions, and prospects. *Applied energy*, 222:1033–1055, 2018.
- [73] Maedeh Ghorbanian, Sarineh Hacopian Dolatabadi, and Pierluigi Siano. Big data issues in smart grids: A survey. *IEEE Systems Journal*, 13(4):4158–4168, 2019.
- [74] Eklas Hossain, Imtiaj Khan, Fuad Un-Noor, Sarder Shazali Sikander, and Md Samiul Haque Sunny. Application of big data and machine learning in smart grid, and associated security concerns: A review. *Ieee Access*, 7:13960–13988, 2019.

- [75] Wooyoung Jung and Farrokh Jazizadeh. Human-in-the-loop hvac operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions. *Applied Energy*, 239:1471–1508, 2019.
- [76] Edward O’Dwyer, Indranil Pan, Salvador Acha, and Nilay Shah. Smart energy systems for sustainable smart cities: Current developments, trends and future directions. *Applied energy*, 237:581–597, 2019.
- [77] Sudeep Tanwar, Qasim Bhatia, Pruthvi Patel, Aparna Kumari, Pradeep Kumar Singh, and Wei-Chiang Hong. Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward. *IEEE Access*, 8:474–488, 2019.
- [78] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, 10(3):3125–3148, 2018.
- [79] Zidong Zhang, Dongxia Zhang, and Robert C Qiu. Deep reinforcement learning for power system applications: An overview. *CSEE Journal of Power and Energy Systems*, 6(1):213–225, 2019.
- [80] Oyenyi Akeem Alimi, Khmaies Ouahada, and Adnan M Abu-Mahfouz. A review of machine learning approaches to power system security and stability. *IEEE Access*, 8:113512–113531, 2020.
- [81] Ibrahim Alotaibi, Mohammed A Abido, Muhammad Khalid, and Andrey V Savkin. A comprehensive review of recent advances in smart grids: A sustainable future with renewable energy resources. *Energies*, 13(23):6269, 2020.
- [82] Ioannis Antonopoulos, Valentin Robu, Benoit Couraud, Desen Kirli, Sonam Norbu, Aristides Kiprakis, David Flynn, Sergio Elizondo-Gonzalez, and Steve Wattam. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, 130:109899, 2020.
- [83] Yassine Himeur, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. Building power consumption datasets: Survey, taxonomy and future directions. *Energy and Buildings*, 227:110404, 2020.
- [84] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wannan Zhang. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831, 2020.
- [85] Martins O Osifeko, Gerhard P Hancke, and Adnan M Abu-Mahfouz. Artificial intelligence techniques for cognitive sensing in future iot: State-of-the-art, potentials, and challenges. *Journal of Sensor and Actuator Networks*, 9(2):21, 2020.
- [86] D Mariano-Hernández, L Hernández-Callejo, A Zorita-Lamadrid, O Duque-Pérez, and F Santos García. A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis. *Journal of Building Engineering*, 33:101692, 2021.
- [87] Muhammad Baqer Mollah, Jun Zhao, Dusit Niyato, Kwok-Yan Lam, Xin Zhang, Amer M. Y. M. Ghias, Leong Hai Koh, and Lei Yang. Blockchain for future smart grid: A comprehensive survey. *IEEE Internet of Things Journal*, 8(1):18–43, 2021.
- [88] Benjamin Völker, Andreas Reinhardt, Anthony Faustine, and Lucas Pereira. Watt’s up at home? smart meter data analytics from a consumer-centric perspective. *Energies*, 14(3):719, 2021.

- [89] Liang Zhang, Jin Wen, Yanfei Li, Jianli Chen, Yunyang Ye, Yangyang Fu, and William Livingood. A review of machine learning in building load prediction. *Applied Energy*, 285:116452, 2021.
- [90] Seungjae Lee and Panagiota Karava. Towards smart buildings with self-tuned indoor thermal environments—a critical review. *Energy and Buildings*, 224:110172, 2020.
- [91] David Connolly, Henrik Lund, and Brian Vad Mathiesen. Smart energy europe: The technical and economic impact of one potential 100% renewable energy scenario for the european union. *Renewable and Sustainable Energy Reviews*, 60:1634–1653, 2016.



Energy utilization of smart or intelligent buildings refers to the definition, modeling, and integration of disparate elements into coherent energy systems in buildings with the help of artificial intelligence. A core aspect of applications for smart building energy is to address the issues of energy utilization directly while simultaneously taking into account user comfort, security and malfunctions. Being deployed in increasing numbers in built environment, smart building energy applications are important components of the built environment today. Given the risen number of renewable energy sources together with tightened regulation to energy consumption, the smart building energy applications provide means to combine new technology components together with heterogeneous requirements and goals for energy utilization in buildings. These goals comprise of, for instance, optimal scheduling of energy consumption and production, optimization of costs, integration of renewable energy, user behavior recognition, and consumer comfort. Overall, machine learning models solve heterogeneous problems in the field of smart building energy utilization. The results of this research indicate that the proposed solutions can provide answers to a variety of issues regarding building energy management, smart grid, personalization, and maintenance and security.



ISBN 978-952-64-1058-6 (printed)  
ISBN 978-952-64-1059-3 (pdf)  
ISSN 1799-4934 (printed)  
ISSN 1799-4942 (pdf)

**Aalto University**  
**School of Science**  
**Department of Computer Science**  
[www.aalto.fi](http://www.aalto.fi)

**BUSINESS +  
ECONOMY**

**ART +  
DESIGN +  
ARCHITECTURE**

**SCIENCE +  
TECHNOLOGY**

**CROSSOVER**

**DOCTORAL  
THESES**