NOVA IMS Information Management School

DOCTORAL PROGRAMME

Information Management

Intelligent Computing Techniques for Clustering and Predicting Energy Consumption in Public Buildings

Ahmed Abdelaziz Mohamed Mohamed

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Intelligent Computing Techniques for Clustering and Predicting Energy Consumption in Public Buildings

by

Ahmed Abdelaziz Mohamed Mohamed

Doctoral Thesis presented as a partial requirement for obtaining the Ph.D. in Information Management

Supervisor: Vitor Duarte dos Santos, Phd, Assistant Professor at Nova IMS

Supervisor*:* Miguel Sales Dias, Phd, Full Professor at Iscte

November 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Ahmed Abdelaziz Mohamed Mohamed]

[Lisbon, November 2023]

Acknowledgments

I want to express my deepest gratitude to my two esteemed professors, Vitor Duarte dos Santos and Miguel Sales Dias, for their invaluable guidance, unwavering support, and insightful feedback throughout this research and writing this thesis. Their expertise and encouragement have been instrumental in shaping the outcome of this work.

I sincerely thank the Adene Agency for generously providing the dataset that served as the foundation for my thesis. Their invaluable contribution significantly enriched the depth and scope of my research, allowing for a comprehensive analysis and meaningful conclusions.

I also want to extend my heartfelt thanks to my family, who have been my pillar of strength throughout this academic journey. Your love, understanding, and constant encouragement have been the driving force behind my perseverance.

In loving memory of my father, Abdelaziz Mohamed Abozeid, whose unwavering belief in my potential inspires me. Although he is no longer with us, his legacy lives on in every achievement and milestone. I am profoundly grateful for the values he instilled in me and for his enduring impact on my academic pursuits.

To all those who supported me in ways big and small, thank you for being part of this journey. Your contributions have left an indelible mark on my personal and academic life.

Abstract

Because intelligent applications may improve the performance of energy consumption, they have recently played a significant role in the energy management of public buildings. Due to their unexpected energy consumption characteristics and the lack of design criteria for sustainable and energy-efficient solutions, these buildings constitute a significant challenge in terms of energy management. Thus, it becomes imperative to investigate the energy usage patterns in public buildings. This highlights how important it is to comprehend and group these buildings' energy usage habits. To assist decision-makers in determining the energy consumption level of each building, this study aims to identify the most intelligent technique for clustering energy consumption of public buildings into levels (e.g., low, medium, and high) and identify critical factors that influence energy consumption. Lastly, predicting energy consumption levels based on clustering model findings utilizing modern intelligence approaches like deep learning techniques.

To achieve the objectives of this study, we proposed three main steps as follows:

First, we put forth two fundamental models: text mining and the PRISMA approach. Using the PRISMA approach, we examined 822 publications between 2013 and 2020 and narrowed the analysis to 106 that satisfied specific criteria, such as having experiments and passing the title and abstract screening stages. The most popular terms and their relationships in the energy and intelligent computing domains were discovered using a text-mining process and a bibliometric map tool (VOS viewer). This allowed researchers to identify the most critical factors influencing building energy consumption and the most effective intelligent computing techniques for grouping and forecasting energy consumption of various building types, particularly public buildings.

Second, two intelligent models, Self-Organizing Map (SOM) and Batch-SOM based on Principal Component Analysis (PCA), were used to determine the number of clusters of energy consumption patterns. We proposed correlation coefficient analysis as a means of identifying critical factors that influence the energy consumption of public buildings. SOM performs better in terms of quantization error than batch-SOM. SOM and Batch-SOM have quantization errors of 8.97 and 9.24, respectively. Two other methods, the Davis-Bouldin method and the Elbow method, were also utilized to calculate the number of clusters. Each building's cluster labels, or levels, were predicted using a genetic algorithm and K-means analysis. In this part, the optimal centroid points in each cluster were identified using a genetic algorithm. If-Then rules have been retrieved by examining cluster levels, so decision-makers must locate the buildings that use the most energy.

Third, Convolutional neural networks (CNNs) and CNNs paired with a Genetic Algorithm (GA) were two intelligent models we suggested using to estimate energy consumption levels. At this stage, we adjusted a few of CNN's settings using a genetic algorithm. The CNN model is beaten by CNN with a genetic algorithm in terms of accuracy and standard error metrics. With accuracy and error of 0.02 and 0.09, respectively, CNN uses a genetic algorithm to achieve 99.01% accuracy on the training dataset and 97.74% accuracy on the validation dataset. On the training dataset, CNN obtains 98.03% accuracy, with 0.05 standard error; on the validation dataset, it achieves 94.91% accuracy and 0.26 standard error.

Finally, this study aids in rationalizing energy usage by building occupants during peak energy consumption periods. It facilitates the replacement of energy suppliers for those buildings by decision-makers in the energy sector. Lastly, we aim to predict energy consumption levels based on clustering model findings utilizing modern intelligence approaches like deep learning techniques.

KEYWORDS

Intelligent Computing Techniques; Clustering; Predictions; Energy Consumption

Sustainable Development Goals (SGD)

Publications

Published articles in Thesis:

Abdelaziz A, Santos V, Dias MS. Machine Learning Techniques in the Energy Consumption of Buildings: A Systematic Literature Review Using Text Mining and Bibliometric Analysis. Energies. 2021; 14(22):7810. https://doi.org/10.3390/en14227810

A. Abdelaziz, V. Santos and M. S. Dias, "Convolutional Neural Network With Genetic Algorithm for Predicting Energy Consumption in Public Buildings," in IEEE Access, vol. 11, pp. 64049-64069, 2023, doi: 10.1109/ACCESS.2023.3284470.

A. Abdelaziz, V. Santos and M. S. Dias, "A Proposed Intelligent Model with Optimization Algorithm for Clustering Energy Consumption in Public Buildings". International Journal of Advanced Computer Science and Applications, 14(9), 136-152. [15]. https://doi.org/10.14569/IJACSA.2023.0140915.

Additional Published articles:

Bahaa A, Abdelaziz A, Sayed A, Elfangary L, Fahmy H. Monitoring Real Time Security Attacks for IoT Systems Using DevSecOps: A Systematic Literature Review. Information. 2021; 12(4):154. https://doi.org/10.3390/info12040154

Abdelaziz A, Anastasiadou M, Castelli M. A Parallel Particle Swarm Optimisation for Selecting Optimal Virtual Machine on Cloud Environment. Applied Sciences. 2020; 10(18):6538. https://doi.org/10.3390/app10186538

Taher, F. & Abdelaziz, A., 2022, "Intelligent Model for Lung Cancer Detection Using Fusion Analysis in CT images", The 3rd International Conference on Distributed Sensing and Intelligent Systems (ICDSIS 2022). Sharjah, United Arab Emirates: Institution of Engineering and Technology, Vol. 2022. p. 252-263 12 p. (IET Conference Proceedings; vol. 2022, no. 14).

Taher, F., & Abdelaziz, A. (2022). Neutrosophic C-Means Clustering with Optimal Machine Learning Enabled Skin Lesion Segmentation and Classification. International Journal of Neutrosophic Science, 19(1), 177-187. https://doi.org/10.54216/IJNS.190113

Alia Nabil Mahmoud, Ahmed Abdelaziz, Vitor Santos, Mario Freire (In Press). A proposed model for detecting defects in software projects, Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), 2023.

Table of Contents

CHAPTER 1 – INTRODUCTION 1
1.1 Scientific Background 1
1.2 Problem Definition 2
1.3 Research Questions 3
1.4 Research Methodology 3
1.5 Research Contribution 4
1.6 Thesis Outline 4
1.7 Path of Research 5
CHAPTER 2 - MACHINE LEARNING TECHNIQUES IN THE ENERGY CONSUMPTION OF BUILDINGS USING TEXT MINING AND BIBLIOMETRIC ANALYSIS 6
2.1 Introduction 6
2.2 Methods 8
2.2.1 Research Questions 10
2.2.2 Search Strategy 10
2.2.3 Text Mining for the Literature 17
2.2.5 Study Selection and Data Extraction 19
2.3 Results and Analysis 19
2.3.1 Text Mining in Detail 19
2.3.2 Bibliometric Map (VOSviewer) 24
2.3.3 Analysis of Representative Manuscripts per Topic 26
2.3.3.1 Analysis of metrics, data sources, and critical factors 26
2.3.3.2 Analysis of clustering and classification techniques 28
2.3.3.3 Analysis of Prediction Techniques 30
2.3.3.4 Analysis of techniques combining classification and prediction 32
2.3.3.5 Analysis of Performance Evaluation Metrics 34
2.4 Discussion 34
2.4.1 Research Question Discussion 34
2.4.2 Research Gap Discussion 37
2.5 Limitations in our Literature Review 38
2.6 Conclusions and Future Work 38
CHAPTER 3 - A PROPOSED INTELLIGENT MODEL WITH OPTIMIZATION ALGORITHM FOR CLUSTERING ENERGY CONSUMPTION IN PUBLIC BUILDINGS 39
3.1 Introduction 39
3.2 Related Work 41
3.3 Research Questions and Methodology 43
3.3.1 Data Collection 44
3.3.2 Data Pre-Processing 46
3.3.3 Feature Selection 49
3.3.4 Finding the Number of Clusters 50

3.3.4.1 Self-Organizing Map 50
3.3.4.2 Elbow Method and Bouldin-Davis Method 52
3.3.5 K-Means with GA 53
3.4 Experimental Results and Discussion 55
3.4.1 Results of Data Pre-processing 55
3.4.2 Results of Feature Selection 58
3.4.3 Results of Finding Number of Clusters 59
3.4.4 KM with GA to Produce Energy Consumption Rules 62
3.5 Conclusion and Future Work 70
CHAPTER 4 - CONVOLUTIONAL NEURAL NETWORK WITH GENETIC ALGORITHM FOR PREDICTING ENERGY CONSUMPTION IN PUBLIC BUILDINGS 72
4.1 Introduction 72
4.2 Related Work 75
4.3 Research Questions and Methodology 78
4.3.1 Dataset and Clustering Preparation 79
4.3.2 CNN with GA 79
4.4 Experimental Results and Discussion 84
4.4.1 Dataset Outputs and Clustering Results 84
4.4.2 CNN with GA to Predict Energy Consumption Levels 84
4.4.3 Implications and practical applications in building energy consumption prediction 87
4.5 Conclusion and Future Work 88
CHAPTER 5 – DISCUSSION 90
5.1 Key Findings 90
5.1.1 Identifying Critical Factors in Energy Consumption 91
5.1.2 Identifying Energy Consumption Patterns 92
5.1.3 Predictive Modeling with Deep Learning 93
5.1.4 Identification of Optimal Settings 94
5.1.5 Potential for Energy Savings 95
5.2 implications and applications 96
5.2.1 Building Operators 96
5.2.2 Energy Suppliers and Grid Operators 96
5.2.3 Policy Makers and Planners 96
5.3 Study limitations 96
5.3.1 Data Availability and Quality 96
5.3.2 Generalizability 96
5.4 discussion summary 97
CHAPTER 6 – CONCLUSION AND FUTURE WORK 98
6.1 Conclusion 98
6.2 Future Work 99
REFERENCES 101

List of Figures

Figure 1. Methodology Steps 9
Figure 2. Search query 11
Figure 3. PRISMA flow chart 11
Figure 4. Word cloud for intelligent techniques applied to energy. 22
Figure 5. Word offset plot for the top 5 words ranked by frequency. 22
Figure 6. General steps for obtaining highly relevant terms in our corpus. The numbers represent the computed *TITs* of the terms. Note: if (*TITs*) > 0.05, the term is relevant 24
Figure 7. The interconnections among the prevailing terminologies as depicted in the bibliometric map. 25
Figure 8. Significant factors of energy consumption in buildings 27
Figure 9. Classification techniques that used energy consumption of buildings 29
Figure 10. Prediction techniques that used energy consumption in buildings 32
Figure 11. Prediction and classification techniques that used energy consumption in buildings. 33
Figure 12. Evaluation Measure of IC Models 34
Figure 13. The most relevant factors that influence the energy consumption of buildings from our survey 35
Figure 14. Top classification techniques identified in our survey. 36
Figure 15. Top prediction techniques identified in our survey. 36
Figure 16. Our Proposed SPKG Model for Discovering Energy Consumption in Public Buildings 44
Figure 17. Structure counts in our data collection, with usage months ranging from 1 to 29. 45
Figure 18. Structure of SOM 51
Figure 19. Data Preprocessing (Stage 1) 56
Figure 20. Data Preprocessing (Stage 2) 57
Figure 21. Degree of polynomials with RSME 57
Figure 22. Total energy use versus contracted power 58
Figure 23. The applied correlation coefficient in the ECD. 58
Figure 24. q- error in Random Weights. Set Iteration = 1000, (a) σ = 0.01, (b) σ = 0.1, (c) σ = 3 and (d) σ = 5 59
Figure 25. q- error in PCA Weights. Set Iteration = 1000, (a) σ = 0.01, (b) σ = 0.1, (c) σ = 3 and (d) σ = 5 60
Figure 26. A Comparison Between PCAW-RTSOM and PCAW-BSOM 61
Figure 27. U-matrix Comparison Between PCAW-RTSOM and PCAW-BSOM 61
Figure 28. Comparison of the Elbow and Davis-Bouldin procedures on a dataset of energy consumption 62
Figure 29. Sample of Clustering Results 63

Figure 30. Monthly energy consumption patterns captured in different clusters for the ECD. 65
Figure 31. Sample of Municipalities that Consume Low Energy Consumption 66
Figure 32. Sample of Municipalities that Consume Medium Energy Consumption 68
Figure 33. Sample of Municipalities that Consume High Energy Consumption 68
Figure 34. Sample of Public Buildings in Different Municipalities that Consume High Energy in 2018 and 2019 69
Figure 35. Our Suggested Model for Categorizing and Estimating Building Energy Use 78
Figure 36. CNN Architecture 80
Figure 37. The Pooling Layer Types 81
Figure 38. The Proposed Flowchart of the GA-enhanced CNN predictive model 81
Figure 39. Acc of CNN Architecture 85
Figure 40. Loss of CNN Architecture 85
Figure 41. Acc of CNN-GA Architecture 85
Figure 42. Loss of CNN-GA Architecture 86
Figure 43. CNN Model Testing and Prediction in Energy Consumption Levels 86
Figure 44. Evaluation of CNN-GA Models for Predicting Energy Use 86
Figure 45. Determine the Common Factors in the Literature Review 91
Figure 46. The Optimal Intelligent Technique to Determine the Best Centroid in the Clusters Regarding SE 92
Figure 47. The Average of Energy Consumption in (2018 and 2019) 93
Figure 48. The Municipalities that Consumed High Energy Consumption in Public Buildings (2018 and 2019) 93
Figure 49. Comparison Between KMGA and State-of-the-Art Methods in terms of SE 94
Figure 50. Comparison Between the Proposed Model (CNN-GA) and State-of-the-Art Methods in terms of MAE 94

List of Tables

Table 1. Studies current stage 5
Table 2. Comprehensive overview of the chapters pertaining to the topic of energy usage in buildings 12
Table 3. Comprehensive overview of the chapters pertaining to the classification of energy use 13
Table 4. Comprehensive overview of the chapters pertaining to the forecast of energy consumption 14
Table 5. comprehensive overview of the integration of classification and prediction techniques for the analysis of energy consumption 16
Table 6. Dictionary for the "energy" domain 18
Table 7. Dictionary for the "IC models" domain 18
Table 8. Top 30 common terms ranked by higher word count on “IC techniques applied to energy”. 21
Table 9. Major factors of energy consumption of buildings 27
Table 10. Major factors of clustering and classification techniques of energy consumption of buildings 30
Table 11. Significant factors of prediction techniques of energy consumption of buildings 32
Table 12. Major combination prediction and classification techniques of energy consumption of buildings 34
Table 13. Dataset Dimensions of Energy Consumption in Public Buildings 45
Table 14. Polynomial Degree Structure - Various Forms 48
Table 15. A comparison Between Random Weights and PCA Weights 59
Table 16. A Comparison Between PCAW-RTSOM and PCAW-BSOM 60
Table 17. GA Parameters 63
Table 18. A Comparison between KMCKI and SPKG in terms of SE and STDEV 64
Table 19. Sample of Energy Consumption Rules 64
Table 20. Sample of Public Buildings That Consume Low Energy in Each Municipality 66
Table 21. Sample of Public Buildings That Consume Medium Energy in Each Municipality 67
Table 22. Sample of Public Buildings That Consume High Energy in Each Municipality 67

List of Abbreviations

ACC&PRE&REC	Accuracy, precision, and recall
AER	Absolute Error
ANN	Artificial Neural Network
BMU	Best Match Unit
CD	Cosine Distance
CNN	Convolutional Neural Network
ECD	Energy Consumption Dataset
ECPB	Energy Consumption of Public Buildings
ED	Euclidean Distance
GA	Genetic Algorithm
IC	Intelligent Computing
KM	K-Means
KMC	K-Means Clustering
LRF	Long-Range Forecasting
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage of Error
MD	Manhattan Distance
MSE	Mean Square Error
NN	Neural Network
PCA	Principal Component Analysis
PR	Chapter Relationship Metric
PSO	Particle Swarm Optimization
q-error	Quantization Error

RMSE	Root Mean Square Error
SER	Standard Error
SOM	Self-Organizing Map
SPKG	SOM, PCA, KM, and GA
SVM	Support Vector Machine
TITs	The Important Terms

Chapter 1 – Introduction

1.1 Scientific Background

Over the past twenty years, the world's growing energy demand has driven the interest in energy efficiency. Notably, in the past five years, interest in the public buildings sector has increased since it represents the third largest energy consumption of European Countries (Banihashemi et al., 2017). Therefore, these countries (mainly Portugal) are trying to discover the reasons that help reduce the ECPB. These countries, including Portugal, strive to reduce energy consumption in public, commercial, and industrial buildings while targeting, at the same time, improving thermal comfort and guiding occupants in public buildings. Mainly, the ECPB represents the third most significant increase in energy consumption (Berardi, U, 2015) and (Berriel et al.,2017). Stakeholders are then particularly interested in understanding the factors that might help reduce energy consumption in these contexts, such as the specific characteristics of buildings, electricity usage, and weather data, amongst other factors. Additionally, stakeholders also seek to understand and influence consumers' energy usage behavior through compulsory legal means, economic subsidies, or communication and publicity means [(Bhattacharjee et al.,2011), (Bogner et al.,2019)].

Thus, to achieve optimization of energy consumption buildings, it is relevant to understand all factors that influence such consumption. Previous studies (Carbonare et al.,2018) and (Chae et al.,2018), including in the Portuguese energy sector (Banihashemi et al., 2017) have been investigated to find possible causes of energy consumption in public buildings and highlighted the following factors:

- Electricity consumption per capita
- Natural gas consumption per capita
- Residential space heating
- Residential space cooling
- Residential appliances
- CO2 emissions by residential
- Residential energy intensity
- weather information and so on.

The energy sector in Portugal (Banihashemi et al., 2017) and (Agência para a Energia, 2018) seeks to find the actual factors that affect the ECPB. There is still a noticeable increase in public energy intensity in the last three years (Banihashemi et al., 2017). It also seeks (Banihashemi et al., 2017) to find an intelligent model capable of clustering energy consumption levels (e.g., low, medium, high, and so on) based on actual factors affecting public buildings' energy consumption. This model also predicts energy consumption levels for the year through clustering. Therefore, the proposed model helps decision-makers in the energy sector as follows:

- Determine the actual factors that affect energy consumption.
- Create an intelligent model capable of clustering and predicting energy consumption in public buildings.
- Guiding people's energy use scientifically and reasonably by controlling consumer behavior in energy consumption through mandatory legal, economic support, or advertising methods.
- Helping the energy sector by striking a balance between actual demand and primary energy demand.

1.2 Problem Definition

Over the past three decades, the efficiency of HVAC stands for heating, ventilation, air conditioning, and other types of equipment has significantly increased in public buildings. Therefore, the energy sector seeks to reduce the amount and slow energy consumption growth in public buildings. Energy consumption in the public building sector represents a high proportion of many countries, accounting for 39% of total energy consumption (Cai et al., 2019). These countries try to provide the energy consumed to maintain their economic, social, and financial level. This study presents a practical case on the state of Portugal and the energy consumption in its public buildings. Energy density in public buildings is increasing at a significant rate of 0.27 in 2018 compared to previous years. Also, the total energy consumption in the public buildings sector, according to the end-use per capita, amounts to 32% of the total energy consumption. Therefore, the energy sector in Portugal (Banihashemi et al., 2017) and (Agência para a Energia, 2018) seeks to conserve energy as much as possible and find the direct factors that lead to high energy consumption and finally make a prediction of energy consumption for years to come to prepare the required amount of energy or predict the levels of energy consumption to determine which public buildings and municipalities that consume high energy consumption to guide their occupants as much as you can.

Obtaining knowledge about the physical, technological, climatic, and behavioral characteristics of a dwelling and its occupants is essential to address the complexity of these factors determining the energy consumption patterns in the public buildings sector. Moreover, this improved knowledge will help (i) better energy planning through more accurate, reliable methodologies encompassing several energy determinants; (ii) to feed targeted-oriented policies, e.g., to specific groups of consumers, such as those under fuel poverty or neighborhoods, and (iii) to be integrated with a boarder framework of policy analysis (i.e., country or city level) evaluating the role of energy policies and instruments for the public buildings sector.

Therefore, there are numerous examples (Banihashemi et al., 2017) and (Agência para a Energia, 2018) of energy efficiency measures that have not always resulted in lower energy consumption, which may partially offset the improvements in efficiency through greater use of equipment or improved comfort (increased energy services demand). Thus, reducing energy consumption also does not mean the need to reduce energy services demand. All these variations result from the interactions between the different energy consumption determinants.

Researchers have suggested many factors to cluster and predict energy consumption in buildings, such as (Delzendeh et al.,2017) and (Diao et al.,2017). However, no formal study has determined the critical factors to help the energy sector cluster and predict energy consumption reasonably. Most researchers also used scientific methods and models to cluster and predict energy consumption, but these models were weak in ACC and results (Djenouri et al.,2019). Moreover, there is a lack of knowledge on the rationale of consumers' behavior regarding electrical energy consumption in Portuguese buildings, and

particularly in public buildings, using characteristics of these buildings, energy consumption (electricity) levels, etc. Thus, the energy sector needs to tackle to problem of the absence in the literature of a formal study to determine the critical factors to build an intelligent model capable of clustering and predicting energy consumption to improve energy efficiency with high and accurate results.

1.3 Research Questions

This study addresses the following main research questions that we believe are able to tackle the identified problem. We have identified two main research questions, stated as follows:

RQ1: How can we build an intelligent model that takes into account the critical factors in public buildings' electrical energy consumption in Portugal (regarding the characteristics of buildings and their electrical power settings), and is able to identify the number of electrical energy consumption clusters with a data science based approach.

RQ2: How can we build an optimized intelligent computing algorithm, for efficiently predicting electrical energy consumption levels in public buildings (based on the clustering model mentioned in the first research question), to enable the public stakeholders to identify the Portuguese buildings that consume high energy, aiming at improving their decision support process regarding energy consumption.

1.4 Research Methodology

The research methodology is composed of three main steps, as follows:

The first step is to review the recent studies of intelligent algorithms in clustering and predicting techniques used in energy consumption models. This step presents the evaluation criteria based on text mining and bibliometric analysis of selected research to determine the energy consumption factors, intelligent techniques, and performance metrics used in the energy sector based on ACC, usability, agility, and applied methods. This step aims to find suitable proposed algorithms to solve the research problem.

In the second step, in the first stage, the proposed model determines the critical factors that influence the ECPB. In the second stage, we used three methods to determine the number of clusters: SOM, Elbow, and Davis-Bouldin. In addition, we built a hybrid intelligent model to predict the clustering labels by KMC with GA.

In the third step, we built a hybrid intelligent model, a GA with a CNN to predict energy consumption levels to help stakeholders identify Portuguese building activities and municipalities that consume high energy consumption.

1.5 Research Contribution

The contribution of this study is divided into three main phases, as follows:

The first phase aims to determine the common factors in the energy consumption of buildings by using recent techniques, which are text mining and bibliometric analysis. The second phase is divided into two stages. The first stage aims to determine the number of clusters using many common techniques extracted from previous works: the SOM, the Elbow method, and the Davis-Bouldin method. The second stage aims to predict clustering labels (levels) energy consumption in each public building in our dataset; therefore, we proposed a hybrid model between optimization algorithms like GA and KMC to fulfill this objective. The third phase aims to predict energy consumption levels of public buildings by using recent intelligent techniques; therefore, we proposed a hybrid intelligent model (GA with CNN) to fulfill this objective.

1.6 Thesis Outline

The following outline describes the content of each of the five sections of the present research, as follows:

The current Section 1 is the introductory section, including the problem statement, scientific background, objectives, and the contribution of this research, as well as the document's outline.

In Section 2, we outline the procedures employed in this study. These methodologies encompass the systematic literature strategy, which incorporates specific inclusion and exclusion criteria for the selection of manuscripts, the utilisation of text mining techniques, and the formulation of research questions. In this study, we undertake the task of describing, analysing, and discussing our findings. Our analysis encompasses a comprehensive examination of text mining processes, bibliometric map analysis using Vosviewer, and an assessment of representative manuscripts pertaining to each topic. In this paper, we outline the limitations of our investigation. In conclusion, the present study encompasses a comprehensive examination of the obtained results. Additionally, it highlights various areas within the research domain that require additional investigation. Based on the findings, conclusive remarks are drawn, and potential avenues for future research are proposed.

In Section 3, we provide an overview of the relevant literature pertaining to current investigations focused on the determination of cluster quantity and the prediction of cluster labels. In this study, we outline our research inquiries and the proposed approach, which centres around the development of an IC model for clustering. Additionally, we provide empirical findings and a comprehensive analysis of the clustering framework. In conclusion, we present our findings and propose avenues for future research.

Section 4 overviews our pertinent research endeavors employing machine learning and deep learning methodologies. The research inquiries and methodologies are delineated. In this study, we proceed to describe and analyse the results obtained from our modelling experiments, specifically focusing on the outcomes derived from the implementation of a hybrid model that combines both the GA and CNN technique. In conclusion, we present our findings and propose suggestions for future research endeavors.

Finally, the paper provides an in-depth analysis of the primary study's topic, presents the conclusions drawn from the research, and offers recommendations for further investigations.

1.7 Path of Research

Our work is a collection of separate research of interrelated subjects, namely IC Techniques for clustering and predicting energy consumption in public buildings, reported separately in different chapters, The current stage of each of the studies is presented in Table 1, as follows:

Table 1. Studies current stage

Chapter	Study name
1	Introduction
2	Machine Learning Techniques in the Energy Consumption of Buildings: A Systematic Literature Review Using Text Mining and Bibliometric Analysis
3	A Proposed Intelligent Model with Optimization Algorithm for Clustering Energy Consumption in Public Buildings
4	Convolutional Neural Network with Genetic Algorithm for Predicting Energy Consumption in Public Buildings
5	Discussion
6	Conclusion and Future Work

Chapter 2 - Machine Learning Techniques in the Energy Consumption of Buildings Using Text Mining and Bibliometric Analysis

Abdelaziz A, Santos V, Dias MS. Machine Learning Techniques in the Energy Consumption of Buildings: A Systematic Literature Review Using Text Mining and Bibliometric Analysis. Energies. 2021; 14(22):7810. https://doi.org/10.3390/en14227810

Buildings exhibit a substantial energy consumption rate, hence exerting a noteworthy influence on the energy efficiency-related behaviours of its occupants. The researchers did a thorough literature analysis in order to determine the key parameters that have the most significant impact on energy usage in buildings. Furthermore, the objective of the review was to identify the most efficient IC techniques that have the capability to anticipate and classify energy usage in various categories of buildings. The study employed the PRISMA methodology to examine a total of 822 scholarly articles that were published from 2013 to 2020. The analysis specifically concentrated on 106 articles, which were selected based on a screening of their titles and abstracts. The study also performed tests and used a text mining methodology with a bibliometric map tool (VOS viewer) to ascertain the frequently employed phrases and their interconnections within the energy and IC fields. The findings of the study indicate that the phrases "consumption," "residential," and "electricity" hold the most significance in the energy domain when considering the ratio of TITs. On the other hand, the term "cluster" emerges as the most frequently utilised keyword in the IC domain. The study additionally revealed a robust correlation between "Residential Energy Consumption" and "Electricity Consumption," as well as between "Heating" and "Climate." In conclusion, the study conducted a comprehensive analysis of 41 articles, employing a critical approach to evaluate their respective contributions. The findings of this analysis facilitated the identification of noteworthy research gaps that warrant additional inquiry.

2.1 Introduction

In recent decades, there has been a notable surge in global energy use, leading to a heightened focus on enhancing energy efficiency within residential, commercial, and governmental infrastructure. The energy consumption of buildings is a significant challenge in numerous European countries due to their status as the primary energy consumers. This is particularly true for public buildings, residential structures, and other public institutions that experience high levels of usage frequency (Nguyen & Aiello, 2013). In the European Union Member States, the transport sector constituted 37% of the overall final energy consumption in the year 2019. This was succeeded by the homes sector, accounting for 32% of the consumption, while the industry and services sectors contributed 42% and 23% respectively. Furthermore, it is worth noting that there has been a significant increase in the efficiency of appliances and equipment in recent years (Kleszcz-Szczyrba, 2010). Hence, European nations, particularly Portugal, are actively endeavouring to enhance the energy efficiency of buildings while simultaneously ensuring adequate thermal comfort and energy consumption levels, with the objective of sustaining their economic and social well-being (Swan & Ugursal, 2009). The energy sector endeavours to manage overall energy consumption through the examination of various data sources and the analysis of different dimensions derived from these sources. These dimensions include, but are not limited to, data on natural gas and electricity usage, characteristics of residential buildings and their energy performance, data on cooling and heating systems, as well as climate and weather forecast data (M. Zhang & Bai, 2018). Public authorities also aim to influence citizen behaviour towards more efficient energy usage by implementing systematic strategies supported by scientific evidence, such as legal regulations, economic incentives, and campaigns to promote best practises (Javaid et al., 2017).

As previously mentioned, investigating energy usage in various sectors of the building stock, such as residential, public services, and industrial, is a pertinent area of research. The existing body of literature has predominantly utilized conventional statistical methodologies for categorizing and forecasting energy usage, yielding outcomes prone to inaccuracies. The absence of effective energy consumption management in many building sectors has resulted in substantial financial losses for numerous nations. Energy stakeholders globally are actively pursuing strategies to decrease energy consumption and enhance efficiency by influencing the behavior of building occupants across many sectors (Jozi, Pinto, Praça, & Vale, 2019). Stakeholders actively seek IC solutions that accurately identify and anticipate energy use, considering the essential elements influencing this consumption.

It is worth mentioning that in Portugal, there is a significant focus within the energy sector on identifying the underlying factors that influence the energy consumption of residential and service buildings. This attention stems from the observed rise in residential energy intensity over the past three years, as reported by Agência para an Energia (2018) and Jozi et al. (2019).

The present research in the field is actively pursuing intelligent models that can classify energy consumption levels (e.g., low, medium, and high levels) (Jozi et al., 2019). These models aim to accurately identify the elements that directly impact energy consumption in residential or service buildings. According to Wood and Amjady (2015), these models also can forecast energy usage in subsequent periods. Intelligent models have the potential to assist various stakeholders within the energy sector, including decision-makers and the general public. These models can serve multiple purposes, such as identifying the key factors that impact energy consumption (Javaid et al., 2017), categorizing and predicting energy usage in residential and commercial buildings, enhancing energy efficiency in these structures, promoting positive behavioral changes among occupants (Massana, Pous, Burgas, Melendez, & Colomer, 2016), and enabling informed decision-making when switching energy suppliers (Qamar & Khosravi, 2015).

Numerous systematic surveys have been documented in the academic literature, which provide comprehensive coverage of pertinent study findings. Our examination of the literature identified a total of eleven review manuscripts that were published within the time frame of 2019 to 2021. According to the study conducted by Runge and Zmeureanu in 2019, Bourdeau et al. (2019) have published scholarly articles that discuss the application of machine learning techniques for forecasting energy consumption in buildings. According to the study conducted by Qolomany et al. (2019), In addition, the study conducted by Djenouri, Laidi, Djenouri, and Balasingham (2019) demonstrated the utilization of machine learning techniques and big data in the context of intelligent buildings. In their respective studies, Vázquez-Canteli and Nagy (2019) and Mason and Grijalva (2019) introduced a novel methodology to enhance building control through machine learning techniques and reinforcement learning. In their study, Guyot et al. (2019) proposed a methodology that utilizes ANN to enhance energy applications across several building sectors. Amasyali and El-Gohary (2018) and Mosavi and Bahmani (2019) demonstrated the application of machine learning methods in predicting energy consumption across different types of buildings. Perera and Kamalaruban (2021) comprehensively analyzed energy system applications using reinforcement learning techniques. However, the review, as mentioned earlier, focused solely on examining a singular factor that impacts the energy consumption of buildings, such as building control, electricity usage, or natural gas consumption. Alternatively, they may have centered their analysis on a specific application, such as occupant behavior or load forecasting.

Additionally, some studies were limited in scope as they solely utilized a particular IC technique, such as ANN and Several recent research has been conducted by Abualigah, Diabat, Mirjalili, Abd Elaziz, and Gandomi (2021), Kacprzyk (2014), Abualigah, Gandomi, et al. (2021), and Abualigah et al. (2020). These studies offer innovative, intelligent approaches that can assist stakeholders in the energy sector

in enhancing energy efficiency. The majority of literature studies in this field exhibit a deficiency in identifying the crucial components that influence the energy consumption of buildings. Moreover, numerous research use conventional statistical or manual techniques to categorize and predict energy consumption.

To bridge these knowledge gaps, our research employed a text-mining method to identify standardized terminology encompassing a wide range of parameters that impact the energy consumption of buildings. This approach was undertaken to address the aforementioned gaps in knowledge effectively. Furthermore, the technology demonstrated an automated and precise identification of the most often utilized intelligent techniques. Moreover, bibliometric analysis was employed to ascertain the interconnections between various aspects and applications of the energy usage of buildings and IC methodologies.

This chapter presents a comprehensive assessment of the existing literature on the application of IC approaches in analyzing and optimizing the energy consumption of buildings. The chapter utilizes the PRISMA methodology (Moher et al., 2009), a straightforward text mining approach, alongside a bibliometric map analysis conducted using Vosviewer (Trianni, Merigó, & Bertoldi, 2018). As mentioned earlier, the approach was employed to ascertain the frequently utilized terminologies within the realms of energy and IC, facilitating the subsequent article analysis phase. This chapter provides an overview of machine learning techniques and other methodologies suitable for clustering and categorizing energy consumption patterns in buildings. These methods apply to ranking low, medium, and high energy consumption levels. Additionally, this paper examines IC techniques utilized in predicting energy usage. Furthermore, this study examines multiple scholarly contributions in the literature, employing a combination of methodologies to accomplish its objectives. This study examines the most viable IC models to classify and predict building energy consumption. Furthermore, this chapter assists scholars in identifying promising avenues for future research within the field.

The primary aims of this chapter are to ascertain the key factors that impact the energy consumption of buildings, determine the prevailing IC techniques employed, forecast and categorize energy consumption in said buildings, and ultimately, identify the performance metrics utilized in the existing literature pertaining to such scenarios.

The chapter is structured in the following manner. In Section 2, we outline the methodology employed in this study. These methodologies encompass the systematic literature strategy, with specific inclusion and exclusion criteria for selecting relevant publications. The text mining method utilized in this research and the research objectives that guided our investigation are also described. In the third section, we provide a comprehensive examination of our findings, which includes an in-depth exploration of text mining techniques, the utilization of Vosviewer for bibliometric map analysis, and an evaluation of representative manuscripts for each topic. In this section, we outline the limitations of our study. An extensive examination of the outcomes is undertaken in the subsequent section, denoted as section 5. Additionally, several areas of inquiry that have not been adequately addressed are identified, leading to the formulation of conclusive remarks. Furthermore, potential avenues for future investigations are proposed.

2.2 Methods

The technique employed in this literature survey is structured into two primary components: In the initial section, we provide a conventional approach for identifying and choosing published publications. In the subsequent section, we elucidate our survey findings by employing text mining and bibliometric analysis techniques, as depicted in Figure 1.

The present study entails a systematic literature survey that assesses the scientific community's contributions regarding energy consumption in buildings. The evaluation uses a rigorous and auditable methodology based on the PRISMA technique.

The PRISMA approach consists of five distinct phases, which are outlined as follows:

1. We are identifying pertinent manuscripts within the specified scope or domains.
2. The process involves the evaluation of titles, abstracts, and chapters, with the exclusion of chapters without empirical evidence and chapters that primarily present positions or opinions.
3. Analysis of eligibility.
4. Full-text screening.
5. The concluding chapters warrant a comprehensive analysis.

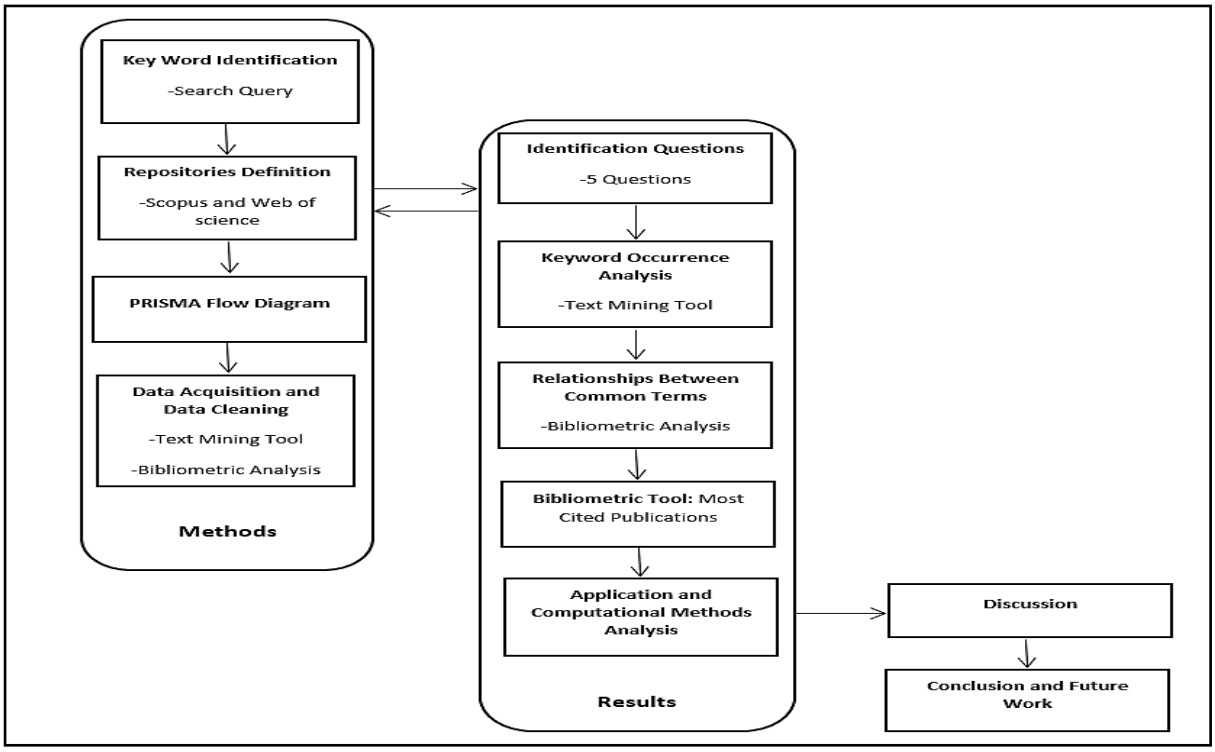


Figure 1. Methodology Steps

In addition, we employed a text-mining methodology and conducted a bibliometric map analysis. The latter approach is employed to establish the connections between commonly used phrases in the energy and machine learning domains (Ma, Wang, & Li, 2020). Text mining techniques identify the most pertinent terms pertaining to the energy and machine learning domains. To achieve this objective, we conducted a study consisting of three distinct phases, during which we assessed the following variables:

1. Word frequency in its whole.
2. The topic of discussion pertains to the words that are frequently encountered or utilized.
3. The frequency of these words is used in the final papers of the study.

This part is organized according to the PRISMA framework, which includes the following components: (1) formulation of research questions, (2) development of a comprehensive search strategy, (3) use of text mining and bibliometric map analysis techniques, (4) establishment of inclusion and exclusion criteria, and (5) final selection of relevant chapters.

2.2.1 Research Questions

The objective of our work is to conduct a comprehensive evaluation of contemporary research endeavors pertaining to the classification and prediction of energy consumption in buildings. The reader is provided with an introduction to the specific themes pertaining to our study objectives and techniques. The present survey focuses explicitly on the following research inquiries, intending to identify the various strategies employed in building energy usage:

- Research Question 1: What measurements, data sources, and crucial aspects have been utilized in previous studies pertaining to the profile of energy consumption in buildings?
- Research Question 2: What are the most effective machine learning algorithms for grouping and classifying building energy usage patterns?
- Research Question 3: What are the most effective machine learning algorithms for accurately estimating buildings' energy usage?
- Research Question 4: What machine learning algorithms provide optimal performance in both the classification and prediction of energy consumption in buildings?
- Research Question 5 (RQ5): Which performance indicators have been utilized in the existing literature to classify or predict the energy consumption of buildings?

In addition, we employed a text-mining methodology and conducted a bibliometric map analysis. Ma et al. (2020) employed the latest techniques to establish connections between conventional energy and machine learning terminologies. Additionally, text mining was utilised to identify the most pertinent phrases within the areas of energy and machine learning. In order to achieve this objective, we conducted a study consisting of three distinct phases, in which we assessed the following variables:

- The frequency of words is used in general.
- The following text examines the most frequently used words.
- The frequency of these often used words in the final papers of the study.

The organisation of this section adheres to the PRISMA guidelines and is organised as follows: (1) formulation of research questions, (2) development of a comprehensive search strategy, (3) use of text mining and bibliometric map analysis techniques, (4) establishment of inclusion and exclusion criteria, and (5) final selection of relevant chapters.

2.2.2 Search Strategy

A literature review often suggests conducting a comprehensive search throughout various journal and conference chapter repositories to see whether similar research has already been accomplished, hence facilitating the identification of potentially pertinent studies. In this study, an exploration was conducted on various electronic chapter repositories. The user accessed and reviewed scholarly publications from different academic databases: IEEE Xplore, Science Direct, Springer, Scopus, and Web of Science. The reviewed manuscripts encompassed technical reports, chapters from scientific conferences, and chapters from scientific journals. The search query was designed to exclusively match the search term within the header section of the documents. In our study, we employed alternative keywords that were logically linked through the use of 'OR' or 'AND' expressions. The search string employed in the electronic repositories indicated is illustrated in Figure 2.

(energy OR consumption OR buildings) AND ("data mining" OR "decision support system" OR "business analytics" OR forecasting OR "modern optimization" OR "machine learning" OR "backpropagation neural network" OR "feedforward neural network" OR "convolution neural network" OR "recurrent neural network" OR "K-mean clustering" OR "hierarchal clustering" OR "artificial intelligence" OR prediction OR predictive)

Figure 2. Search query

Figure 3 presents the PRISMA flow chart, which visually represents the five distinct phases employed while filtering the text set. The variable "n" represents the number of chapters obtained after each Phase or Step.

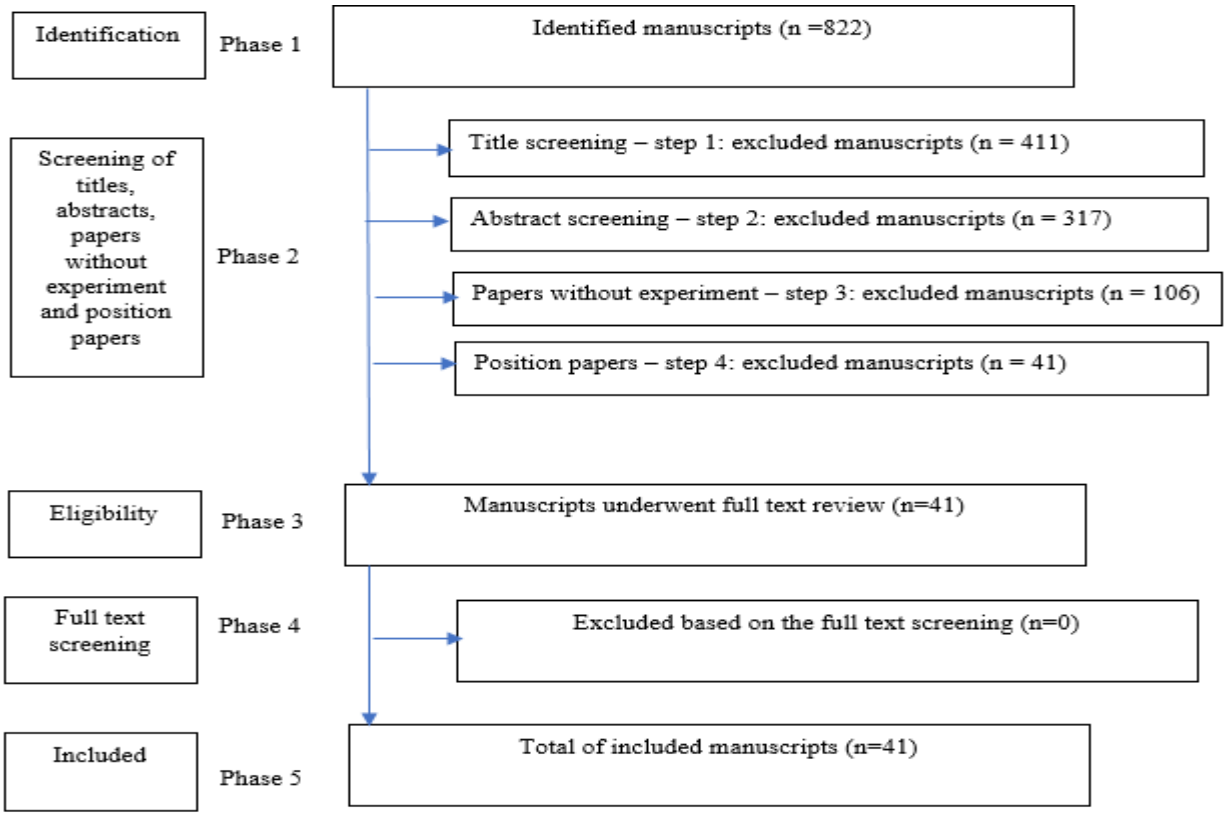


Figure 3. PRISMA flow chart

We conducted a comprehensive search across several electronic archives during the initial phase, utilizing a specific search term. We aimed to identify chapters that were published throughout the timeframe of 2013 to 2020. This search yielded a total of 822 publications. Phase 2 was conducted using a systematic five-step methodology. During the initial stage, we worked an exclusion process whereby manuscripts were eliminated based on their names. For instance, publications pertaining to energy consumption in industry buildings, transport, and services were deleted. As a result, the total number of publications was reduced to 411. During the second step, a total of 317 articles were removed based on the screening of their abstracts. During the third step, 106 publications were identified after excluding manuscripts that reported research without experiments. Following this, during the fourth stage of phase two, we eliminated papers irrelevant to the topic, resulting in a final count of 41 publications. During phase 3 of the study, the manuscripts were subjected to a comprehensive examination and evaluation of their whole text. This rigorous process ultimately resulted in no exclusions, as observed in phase 4 of the study.

The chapter selection approach yielded a final list comprising 41 manuscripts (phase 5), which this chapter will thoroughly examine. These were subsequently categorized into four distinct groups, as seen in Tables 2, 3, 4, and 5.

- The topic of discussion is the energy usage of buildings, specifically buildings labeled as S1 to S10.
- The categorization is used of energy usage in buildings within the range of S11 to S20.
- The objective of this study is to forecast the energy consumption of buildings, specifically in the range of S21 to S33.
- The research integrates categorization and prediction techniques to analyze and forecast building energy consumption patterns, specifically within S34 to S41.

Table 2. comprehensive overview of the chapters pertaining to the topic of energy usage in buildings.

Chapter	Reference	Application	Data dimensions	Method	No. of Citations
S1	(Guerra Santin, 2013)	Occupant behavior in energy-efficient dwellings	Buildings characteristics and occupant behavior	Statistical analysis	80
S2	(Delzendeh, Wu, Lee, & Zhou, 2017)	The impact of occupants' behaviors on building energy analysis	Occupant behavior	Systematic review	195
S3	(Berardi, 2015)	Comparative study between the energy consumption of residential buildings in US, EU, and BRIC countries	Climate, space heating, space cooling, and hot water	Statistical analysis	72
S4	(Mancini, Basso, & De Santoli, 2019)	Energy use in residential buildings: characterization via identifying flexible loads employing a survey questionnaire	Building location, number of occupants in the dwelling, building services system, kitchen	Questionnaires	12
S5	(Csoknyai, Legardeur, Akle, & Horváth, 2019)	Analysis of energy consumption profiles in residential buildings	Electric consumption, heating and hot water	Questionnaires and serious game	14

S6	(Mardookhy, Sawhney, Ji, Zhu, & Zhou, 2014)	A study of energy efficiency in residential buildings	HVAC, electricity, natural gas, and lighting system	Statistical analysis and questionnaires	66
S7	(Cao, Dai, & Liu, 2016)	Building energy-consumption status worldwide and state-of-the-art technologies for near zero-energy buildings	Climate change, space heating, and space cooling	Questionnaires	358
S8	(Chang, Zhu, Yang, & Yang, 2018)	Reduction of energy consumption in residential buildings	Climate information	Statistical analysis	23
S9	(Hannan et al., 2018)	Identification of elements to control and regulate residential energy consumption	Demographics, consumer attitude, economy, and climate	Correlation coefficients	60
S10	(Bhattacharjee & Reichard, 2011)	Socio-economic factors affecting individual household energy consumption	Socio-economic factors	Systematic review	33

Table 3. comprehensive overview of the chapters pertaining to the classification of energy use

Chapter	Reference	Application	Data dimensions	Method	No. of Citations
S11	(Gouveia & Seixas, 2016)	Unraveling electricity consumption profiles in households through clusters: combining smart meters and door-to-door surveys	Electricity consumption and weather information	Hierarchical clustering and door to door surveys	83
S12	(Azaza & Wallin, 2017)	Smart meters data clustering to find the most responsible consumers in the peak system	Responsibility factor and consumption variability	Hierarchical clustering and self-organizing map	15
S13	(Gouveia, Seixas, & Mestre, 2017)	Daily electricity consumption profiles from smart meters	Climate, space heating, space cooling and electricity consumption	Hierarchical clustering and door to door surveys	28
S14	(Diao, Sun, Chen, & Chen, 2017)	Discovering electricity consumption over time for residential consumers	Occupant behavior	K-means clustering	72

S15	(Nepal, Yamaha, Sahashi, & Yokoe, 2019)	Analysis of building electricity use pattern	Human activities and air conditioning	K-means clustering	14
S16	(Jin et al., 2017)	Comparison of clustering techniques for residential energy behavior using smart meter data	Smart meter data	Hierarchal clustering and K-means clustering	25
S17	(Carbonare, Pflug, & Wagner, 2018)	Clustering the occupant behavior in residential buildings	Window opening and indoor temperature	K-means clustering and time series	39
S18	(Pan et al., 2017)	Cluster analysis for occupant behavior-based electricity load patterns in buildings	Electricity profiles	K-means clustering	17
S19	(Diao et al., 2017)	Modeling energy consumption in residential buildings	Space heating, refrigeration, and air-conditioning	K-means clustering and demographic-based probability neural networks	72
S20	(Tureczek, Nielsen, & Madsen, 2018)	Electricity consumption clustering using smart meter data	Smart meter data	K-means clustering with time series analysis and wavelets	23

Table 4. comprehensive overview of the chapters pertaining to the forecast of energy consumption

Chapter	Reference	Application	Data dimensions	Method	No. of Citations
S21	(Braun, Altan, & Beck, 2014)	Predicting the future energy consumption of a supermarket in the UK	Climate information, electricity consumption, natural gas consumption	Multiple linear regression	172
S22	(Wahid, Ghazali, Shah, & Fayaz, 2017)	Prediction of energy consumption in residential buildings	Occupant behavior	Multilayer perceptron and random forest	16
S23	(Liu et al., 2019)	Machine learning model for forecasting energy consumption of buildings	Weather condition and building envelope	Artificial neural network and support vector machine	18

S24	(T. Y. Kim & Cho, 2019)	Predicting residential energy consumption	Individual household power consumption	Deep neural network	67
S25	(Jozi, Pinto, Praça, & Vale, 2019)	Decision support application for energy consumption forecasting	Total energy consumption and environmental temperature	Neuro-fuzzy algorithm	15
S26	(Moretti, Nassuato, & Bordoni, 2019)	Development of regression models to predict energy consumption in manufacturing companies	Mean outdoor temperature and electricity data	Multiple linear regression	12
S27	(Berriel, Lopes, Rodrigues, Varejao, & Oliveira-Santos, 2017)	Monthly energy consumption forecast: a deep learning approach	Space heating and cooling, class of the customer, and average of the customer consumption	Deep neural network	41
S28	(Edwards, New, & Parker, 2012)	Predicting future hourly residential electrical consumption	Billing electricity data	Regression model, feed-forward neural network, and support vector regression	262
S29	(Rahman, Selvarasan, & Jahitha Begum, 2018)	Short-term forecasting of total energy consumption	GDP, population, and GDP per capita	Multiple linear regression and simple regression model	15
S30	(Wang, Wang, & Wang, 2017)	Influencing factors regression analysis of heating energy consumption of rural buildings	Family basic information, rural residential building features, building envelope information, indoor air quality in winter, and building heating energy consumption	Multiple linear regression and logistic regression	13
S31	(Zekić-Sušac, Mitrović, & Has, 2020)	Machine learning-based system for managing the energy efficiency of the public sector as an approach towards smart cities	Geospatial attributes, construction attributes, heating attributes, and temperature attributes.	Convolution neural network with coefficient correlation, decision tree, and random forest	18

S32	(S. Kim, Jung, & Baek, 2019)	Predicting energy consumption of buildings	Electricity billing data of occupants	Support vector machine	13
S33	(Bogner, Pappenberger, & Zappa, 2019)	Predicting energy consumption in public buildings	Space heating and cooling and weather information	Multivariate adaptive regression splines, quantile regression, quantile random forest, gradient boosting machines, and nonhomogeneous gaussian regression	17

Table 5. comprehensive overview of the integration of classification and prediction techniques for the analysis of energy consumption

Chapter	Reference	Application	Data dimensions	Method	No. of Citations
S34	(Aqlan, Ahmed, Srihari, & Khasawneh, 2014)	A hybrid approach to assess the energy efficiency of residential buildings	Space heating and space cooling	K-means clustering and artificial neural network	14
S35	(Jovanović & Sretenović, 2017)	A hybrid approach to predict heating energy consumption	Individual heating energy consumption	Radial basis neural networks and K-means clustering	12
S36	(Banihashemi, Ding, & Wang, 2017)	Developing a hybrid approach of prediction and classification algorithms for building energy consumption	Building envelops, building design layout	Artificial neural network and decision tree	23
S37	(Seyedzadeh, Rahimian, Glesk, & Roper, 2018)	Propose a machine learning approach for the estimation of building energy consumption.	Amount of gas emission and CO2 emission	Artificial neural network and K-means clustering	53
S38	(Zekić-Sušac, Scitovski, & Has, 2018)	Prediction of energy efficiency of public buildings	Geospatial, construction geometry	Artificial neural network and K-means clustering	16

S39	(Tang, Lee, Wang, & Yang, 2019)	Leveraging socio-economic information and deep learning for residential load pattern prediction	Socio-economic factors	Deep neural network and K-means clustering	16
S40	(Cai, Shen, Lin, Li, & Xiao, 2019)	Presented a novel approach to predict electricity consumption in residential buildings	Building characteristics and weather information	K-means clustering and support vector machine	20
S41	(Gajowniczek & Zabkowski, 2018)	Simulation study on clustering approaches for short-term electricity forecasting	Electricity profiles	K-means clustering, neural network, and support vector regression	14

2.2.3 Text Mining for the Literature

The current investigation utilised the PRISMA methodology to examine a comprehensive collection of 822 scholarly articles that were released between the timeframe of 2013 to 2020. A total of 106 papers were chosen for inclusion in our analysis, with selection criteria based on the examination of their title, abstract, and manuscripts that did not involve experimental procedures. In light of the substantial volume of publications, a text-mining methodology was utilised to examine the pertinent terminologies within the domains of IC models and energy consumption. The objective of this method is to organise data in a manner that enhances the subsequent examination of written works. To fulfil this objective, we have constructed two lexicons, one pertaining to "IC models" and the other concerning "energy consumption of buildings." Each lexicon comprises a compilation of word or phrase expressions. In order to optimise the efficacy of this methodology, we have established a comprehensive lexicon encompassing prevalent terminology within the field, as well as terms specifically linked to the concepts pertinent to our research subject. This strategy exhibits a higher level of comprehensiveness compared to conventional text mining techniques that employ random word searching, grouping, and counting. Therefore, this study showcases the effectiveness of text-mining in the examination of extensive datasets, offering a more organised and efficient methodology for data analysis.

It is noteworthy to mention that each of the four authors demonstrates substantial competence in the areas discussed within the chapter, with a particular emphasis on computer science and machine learning for the first, second, and third authors, and experience in energy for the third author. In order to authenticate the dictionaries, a thorough examination was conducted on the texts, focusing on their titles, abstracts, and keywords. In light of the extensive collection of manuscripts at our disposal, a suitable quantity of articles was chosen at random for the purpose of validating the dictionaries. Tables 6 and 7 present the dictionaries in question. It is important to acknowledge that the terms "energy" and "intelligence" were considered excessively wide and, as a result, were excluded from the dictionaries. Likewise, in instances where a subject matter lies beyond the purview of our investigation, such as "industrial building," it is omitted from our lexicons.

Table 6. Dictionary for the "energy" domain

Nr	Reduced Term	Similar Term
1	consumption	reduce, minimize
2	buildings	constructing, structure
3	occupant	Résidente, inhabités, habitant, consumer
4	behavior	behavior, conduct, attitude, action
5	electricity	electro
6	residential	domestic, household, home
7	public	general, generic, common
8	commercial	mercantile
9	patterns	sample, type, modality
10	heating	warming, hot, heat
11	cooling	refrigeration, cool
12	water	hot water, cool water
13	climate	weather
14	gas	natural gas

Table 7. Dictionary for the "IC models" domain

Nr	Reduced Term	Similar Term
1	artificial intelligence	machine learning, intelligent
2	predict	prediction, predictive, predicting, forecasting, forecast
3	classification	classifier, classifiers
4	cluster	clusters, clustering, K-means cluster, hierarchal
5	model	paradigm, sample
6	method	process, procedure
7	analysis	analytics, data sciences, data science
8	efficiency	performance, quality
9	neural network	neural networks, feedforward, backpropagation, convolution, recurrent
10	regression	time series, linear, logistic
11	decision tree	decision trees, random forests, random forest
12	optimization	optimize
13	approach	approaches
14	study	survey, experiment

During the course of our investigation, it was observed that certain terms from the dictionary were absent from the title, abstract, or keywords of the manuscripts that were examined. This phenomenon occurs when specific terms are reiterated several times inside an article. Consequently, we took into account the entirety of the text during our analysis of the gathered material, with a specific focus on the domain of IC models and the energy consumption of buildings. In order to maintain uniformity, the reference section was excluded from all papers during the course of our research.

In the second part of our terminology analysis, we utilized a bibliometric map to establish significant relationships between factors and IC methods. This bibliometric map facilitates the identification of the most frequently used factors and methods, as well as their relationships, which can be useful for stakeholders.2.2.4 Inclusion and Exclusion Criteria

Our chapter's analysis was conducted on a set of manuscripts that met specific inclusion criteria. These criteria comprised manuscripts that directly addressed one or more of our research questions and were published between 2013 and 2020.

Several criteria were used to exclude manuscripts from our analysis. These included chapters that lacked experimental analysis and those that consisted of viewpoints, books, workshops, or tutorials. Additionally, position chapters were also excluded from our review.

2.2.5 Study Selection and Data Extraction

Upon applying the predetermined inclusion and exclusion criteria, our search of the chapter repository yielded multiple chapters that we thoroughly analyzed and reviewed. Our primary focus throughout this search was to identify any existing scientific research gaps. To facilitate this process, we devised a data extraction form that allowed us to gather pertinent information from the selected primary chapters, thereby enabling us to address our proposed research questions with greater ACC and precision.

2.3 Results and Analysis

The results and analysis have been organised into three sub-sections. The initial section of our study presents the methodology and outcomes of our text mining techniques, encompassing the use of a word frequency table, a word offset plot, and a word cloud plot. The word cloud plot depicts the most frequent terms detected, with the font size indicating their relative frequency (see to Figure 4). The term "offsets plot" is utilised to quantify the dispersion of words within a corpus, as depicted in Figure 5. Through the process of evaluating these values, it becomes possible to visually represent the relative significance of individual words within our corpus. In the subsequent part, a comprehensive examination of the bibliometric map is conducted in order to identify the crucial associations between variables and the prevalent employment of IC approaches in the context of energy consumption within buildings. Finally, inside the third sub-section, an analysis is conducted on the 41 papers that were kept.

2.3.1 Text Mining in Detail

The text mining processes involved a series of pre-processing steps that were applied to the primary documents. The initial step involved the removal of symbols, numerals, punctuations, and whitespaces, followed by the conversion of all words to lowercase. Additionally, we compiled a consolidated inventory of dependable phrases from the dictionary, which we subsequently employed during the process of reading and analysing the most relevant articles for our research. The procedure was implemented utilising algorithm 1, which outlines the essential steps involved in developing a

corpus tool capable of determining the frequency of words in manuscripts. In order to construct our corpus, we initially consolidated all manuscripts into a unified "Documents" file, along with two "Dictionaries" (line 1). Between lines 3 and 10, a series of programming libraries were imported and subsequently run in a predetermined sequence. This facilitated the construction of our corpus in a smooth manner.

- Line 3 (import **re**): we imported a library for regular expression operations.
- Line 4 (import **nltk**): we imported a toolkit for natural language processing.
- Line 5: we computed **SW** = Stop Words.
- Line 6: we computed **£** = the word frequency for all words in **α** ("Documents").
- Line 7 (from **nltk**. **corpus** import **SW**): We computed **corpus** = a large and structured set of texts and SWs.
- Lines 8 and 9: We removed the morphological affixes of words from the **corpus** leaving only the word stem.
- Line 10: We recorded the frequency of each word in **α** ("Documents").

We continued by implementing a cleaning process by deleting from **α** symbols, numbers, and extra spaces (line 11) and transforming all words to lowercase (line 12). Line 13 converted sentences to separate words. Line 14 removed stop words such as "the" and "is" and those that use stemming. That is, we reduced variations of the form of a given word by deleting inflection forms through the removal of unnecessary characters, such as in the following example:

$$\begin{cases} Reduces\ energy \\ Reduced\ energy \\ Reducing\ energy \end{cases} \longrightarrow Reduce\ energy.$$

Line 15 and Line 16 found common words between the main document file and the two dictionaries. Line 17 computed the word frequency in the main document file. Line 18 created a mapping between word frequency and the intersection words (that are common across "Document" and "Dictionary") to find optimal keywords that were used in selecting significant research chapters for our survey. Finally, in lines 19, 20, we determined the importance of each word, found optimal keywords, and visualized a word cloud (Figure 4) and a word offset (Figure 5). In this last one, we depicted the location of a word in a sequence of text sentences.

Algorithm 1: Build a corpus module to find word frequency in manuscripts

1. **Input: α = (Documents** number of manuscripts**)**
   ¥ = (Dictionary two dictionaries of intelligent model and energy consumption of buildings**)**
2. **Output: β ⟵ (Results** Words frequency, importance of each word and word offsets & word cloud visualization**)**
3. **Import re**
4. **Import nltk.**
5. **SW = stopwords**
6. **£ = common word**
7. **from nltk. corpus import SW.**
8. **PS = PorterStemmer ()**
9. **from nltk. stem. porter import PS.**
10. **from nltk. probability import FreqDist.**
11. **review =** re. sub ('[^a-z A-Z]', ' ', α)
12. **review =** review. Lower ()
13. **review =** review. Split ()
14. **review =** [PS. stem(word) **For** word **in** review **if not** word in **set** (stopwords. words ('english'))
15. **def Intersection (review, ¥):**
16. **return set(review). intersection (¥)**
17. **fdist = FreqDist(review)**
18. **Mapping between fdist and intersection**
19. **Implement dispersion plot.**
20. **Determine the importance of each word:**
21. **Return β**

After implementing our text mining methods, a total of 1077 commonly used terms were found from two dictionaries. A selection of these terms is provided in tables 6 and 7. The results of our study have identified the 30 most often occurring terms, as ranked by word frequency, as illustrated in algorithm 1. Among these terms, the first and second terms, namely "consumption" and "buildings," are linked to the domain of "energy," while the third term, "predict," pertains to the realm of machine learning or "intelligent" systems.

We were able to leverage the final list of word frequencies to identify common terms in the "energy" and "IC models" domains. Furthermore, we created word offsets and word cloud plots to facilitate a visual interpretation of the results obtained.

To gain a more detailed comprehension, we conducted an analysis of Table 8, wherein we discovered a correlation between several phrases pertaining to energy consumption and the use of intelligent methodologies by multiple writers to tackle issues within the energy sector. The present analysis offers a comprehensive examination of the variables that influence energy efficiency or consumption in diverse building structures. Additionally, this paper elucidates some IC techniques employed to efficiently tackle these challenges.

The results of our study indicate that contemporary research endeavours are mostly concentrated on the utilisation of advanced intelligent techniques, such as deep learning, in order to successfully tackle energy-related challenges. In summary, the findings of our study offer significant contributions to the understanding of prevalent terminology and intelligent approaches employed within the energy sector. These insights have the potential to guide and shape future investigations and advancements in this particular field of study.

Figure 4 displays a word cloud plot that showcases the terms that are used most frequently. In contrast, Figure 5 exhibits a word offset plot that ranks the top five words based on their frequency. The plot illustrates the sequential placement of a word inside a given text, commencing from its initial occurrence, and can be effectively represented by a dispersion plot. In the dispersion map, each stripe corresponds to an occurrence of a word, while each row represents the entirety of the text. The plots presented in Figures 4 and 5 were generated following the implementation of our stemming technique. This process led to the truncation of numerous words.

Table 8. List of the top 30 standard phrases, ordered in descending order based on their word count, about applying IC techniques in energy.

Rank	Term	Count	Rank	Term	Count
1	consumption	1881	16	heat	363
2	buildings	1639	17	machine	341
3	predict	957	18	chapter	341
4	residential	825	19	result	330
5	cluster	671	20	network	319
6	model	605	21	load	308
7	electricity	550	22	behaviour	308
8	method	495	23	different	308
9	analysis	484	24	neural	308
10	base	451	25	perform	297
11	efficiency	440	26	factor	286
12	occupant	418	27	pattern	286
13	forecast	407	28	regression	286
14	study	396	29	learning	286
15	research	385	30	approach	264

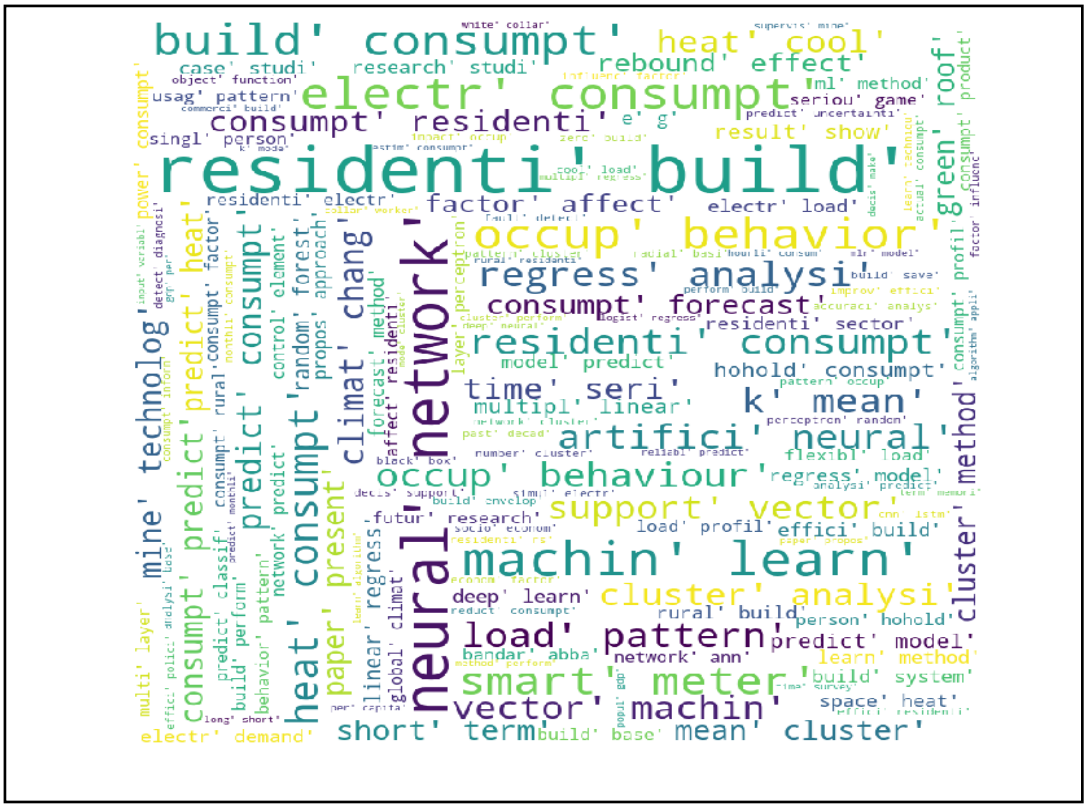


Figure 4. Word cloud for intelligent techniques applied to energy.

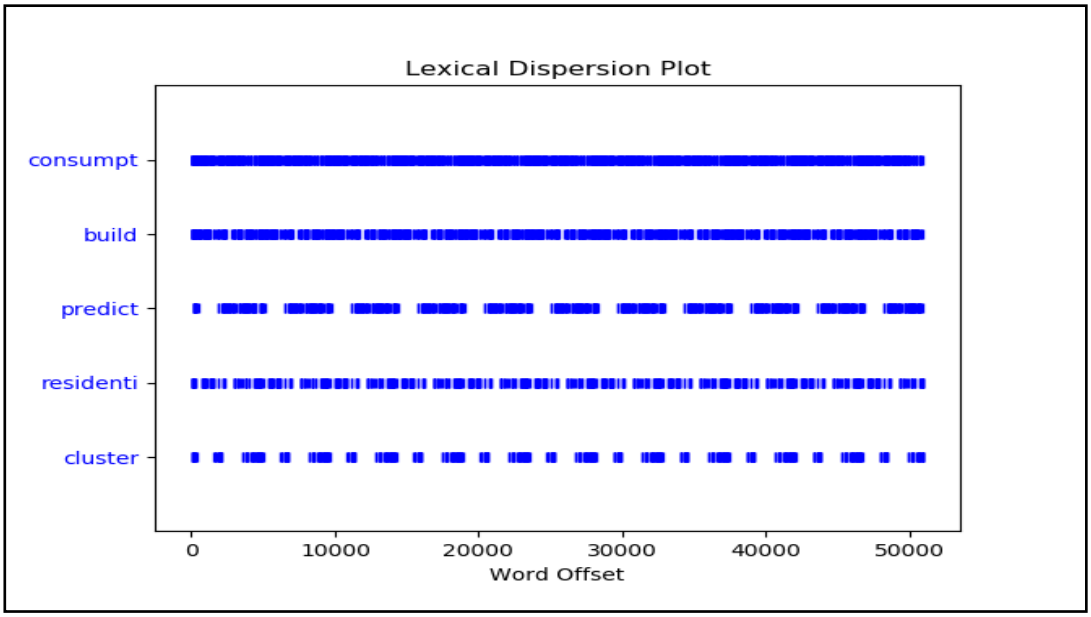


Figure 5. Word offset plot for the top 5 words ranked by frequency.

Figure 6 illustrates the methods utilised to find highly pertinent phrases related to IC strategies implemented in the energy sector. To obtain the frequency of top words relevant to our research, we calculated the intersection between the words used in all manuscripts and two dictionaries of the machine learning and energy domains. To achieve this, we identified a set of keywords, including "intelligent," "method," "energy," and "buildings," which we considered most pertinent to our study. Subsequently, we conducted text mining to identify additional crucial terms, such as "prediction,"

"model," "consumption," and "residential." For each of these, we calculated TITs (Elgendy, Zhang, He, Gupta, & Abd El-Latif, 2021) and (Wu, Liu, Ahmed, Peng, & El-Latif, 2020), which represent the importance of each word in the corpus. The formula for computing TITs is defined in Equation 1.

$$TITs = \frac{\text{review.count}(£)}{\text{len (review)}} \qquad (1)$$

Where:

- *review* = words in our studied corpus
- £ = a common word, such as "consumption" or "cluster"

In the realm of text mining, identifying highly relevant terms is crucial for extracting meaningful insights from large corpuses. To achieve this, researchers use the TITs metric. TITs is calculated as the ratio of the term count to the total count of words in the corpus. A word is considered highly relevant if its TITs value is greater than 0.05. This threshold is based on the percentile value of a term that is considered highly relevant in the corpus, such as "consumption" with a TITs value of 0.31. Conversely, words with TITs values below 0.05, such as "classification," "water," "commercial," and "services," are considered less relevant.

Moreover, the relevance of the identified terms is closely related to the entire corpus. To measure the relationship between the corpus and the identified terms, we employed the PR. PR is calculated using formula 2, as proposed by Ahmed A. Abd El-Latif et al. in 2021. The PR metric provides a quantitative measure of the correlation between the identified terms and the entire corpus.

In summary, the TITs metric is a useful tool for identifying highly relevant terms in the corpus while the PR metric provides an accurate measure of the relationship between these terms and the corpus. These metrics can help researchers extract meaningful insights from large corpuses and enable them to make data-driven decisions.

$$PR = \frac{PCCT}{MRW} 100 \qquad (2)$$

Where:

- *PCCT* = number of chapters that contain standard terms
- *MRW* = all manuscripts in related work

The current investigation examined a collection of written documents in the domain of IC approaches as they pertain to energy. The objective was to ascertain the most significant terms and their frequency of occurrence across the collection. The results of our study indicate a high PR ratio of 41 for the word "consumption of buildings" compared to all manuscripts, representing 70.7% of the corpus. The remaining PR ratio of 29.3% is attributed to the term "efficiency of buildings." Furthermore, the PR ratio reaches its peak when comparing the regular expression "cluster * buildings" with all the manuscripts found in Table 3 and Table 5. These tables pertain to the classification of energy consumption in buildings and collectively represent 100% of the corpus.

Furthermore, an assessment was conducted on the PR ratio pertaining to the terms "neural * buildings" in relation to all manuscripts included in Table 4 and Table 5. These tables specifically pertain to the prediction of energy consumption in buildings. The analysis revealed a PR ratio of 52.4%. The remaining proportion, accounting for 47.6%, is associated with alternative machine learning methodologies that specifically involve the concept of "buildings." In summary, the outcomes of our text mining analysis facilitated the identification of frequently utilised terminology within the domain and facilitated the discovery of the most pertinent scholarly articles pertaining to the subject matter.

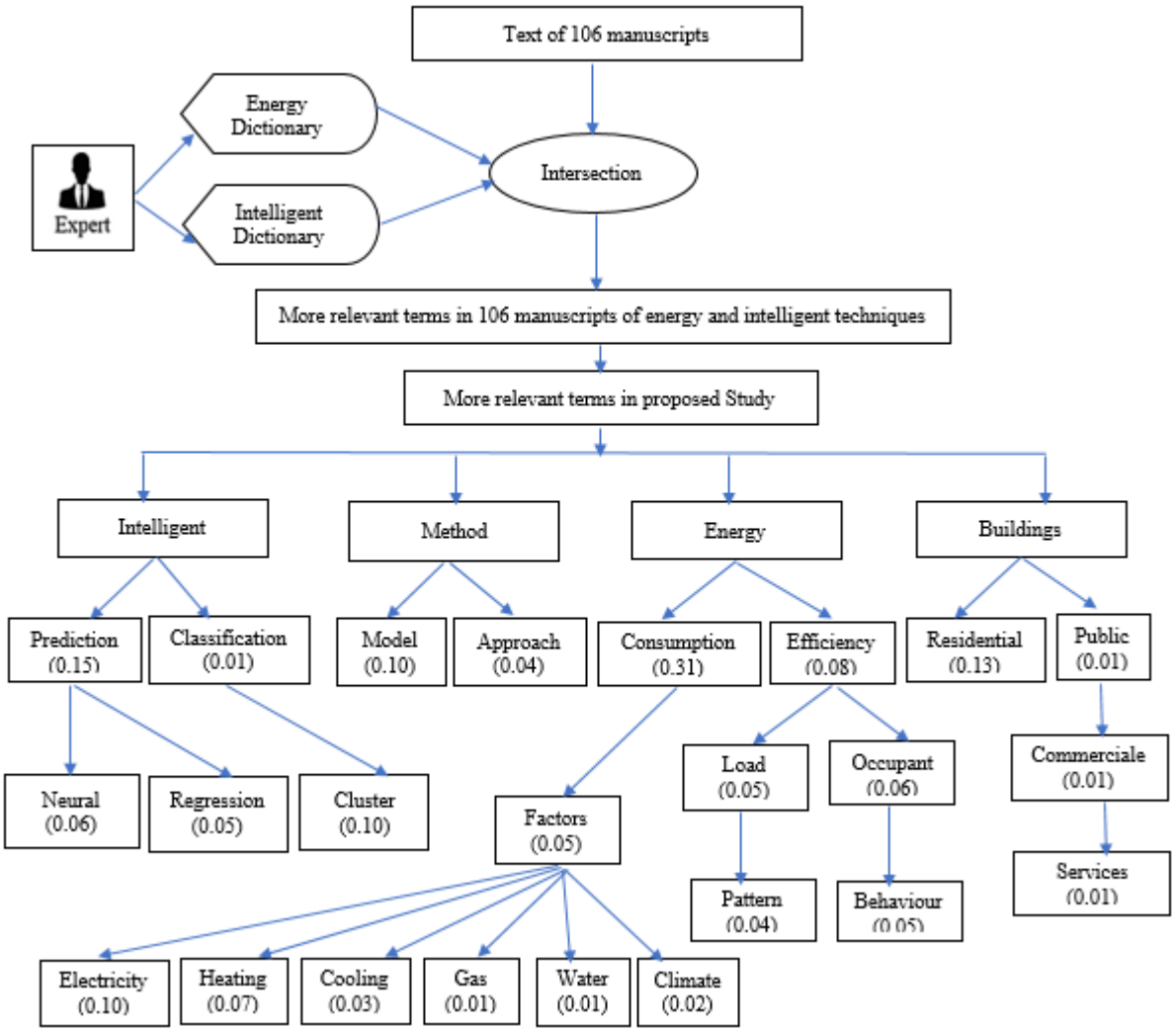


Figure 6. General steps for obtaining highly relevant terms in our corpus. The numbers represent the computed *TITs* of the terms. Note: if (*TITs*) > 0.05, the term is relevant

2.3.2 Bibliometric Map (VOSviewer)

In our study encompassing 41 publications focused on the relationship between energy consumption and machine learning approaches, we employed VOS viewer, a bibliometric network visualisation tool, for our research. The utilisation of this instrument facilitated our research by providing visual data, allowing us to analyse the interconnections between the fields of energy and intelligent techniques. Furthermore, this study facilitated the identification of prevalent dimensions, clustering methods, and methodologies that proved vital in efficiently addressing our research inquiries.

Figure 7 illustrates the network map visualization that displays the relations between the most popular terminologies and how they are linked. The size of the nodes represents the frequency of appearance in the manuscripts. VOS viewer employs a clustering method to group related terminologies according to their relevance to each other.

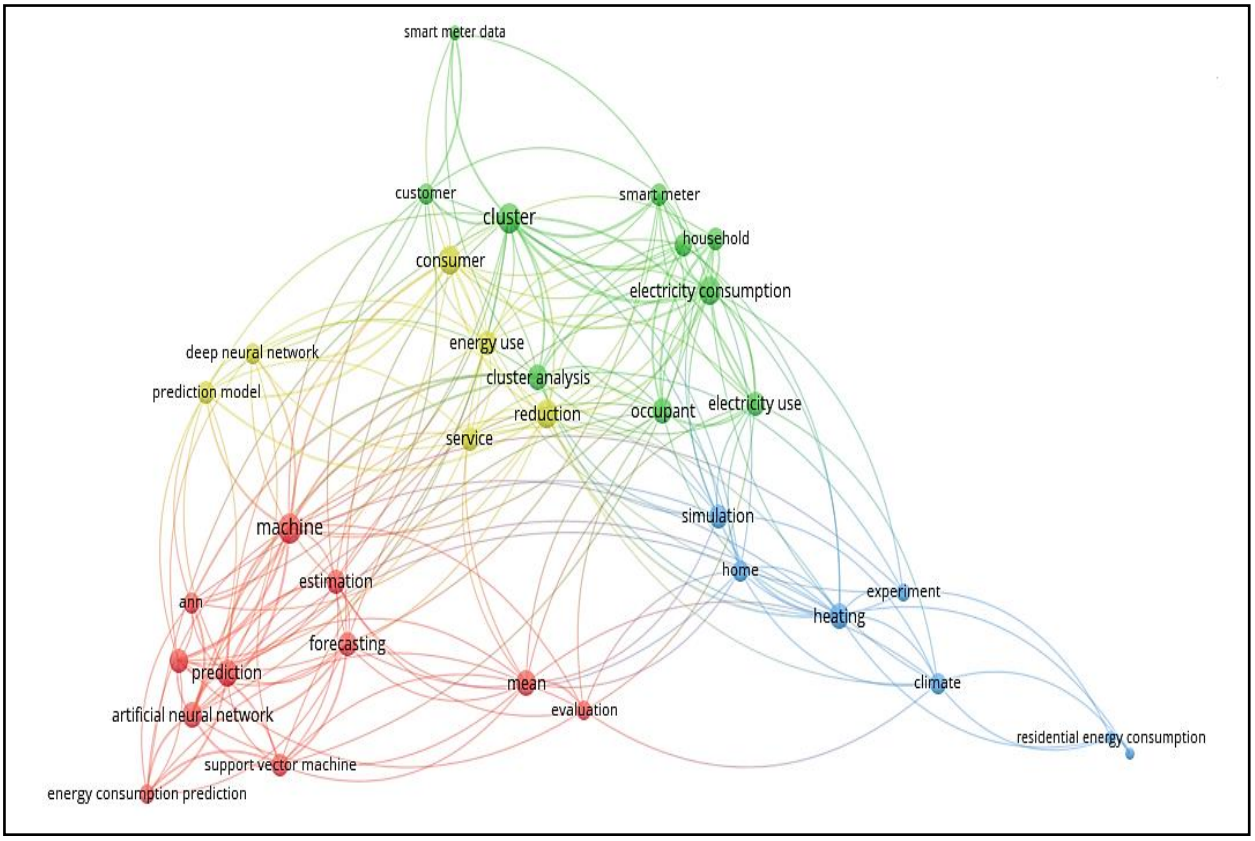


Figure 7. The interconnections among the prevailing terminologies as depicted in the bibliometric map.

The present study conducted an analysis of the title and abstract of a scholarly article using a binary counting method of 1177 examined keywords with a minimum threshold of 3 occurrences. The results of the analysis yielded 33 terminologies, as highlighted in figure 7. The network map of the study revealed the four clusters of terminologies, with the most significant nodes representing the critical nodes of each cluster. The clusters were determined as "NN" and "Energy Consumption Prediction" (red), "Deep NN" and "Energy Use" (yellow), "Cluster" and "Electricity Consumption" (green), and finally, "Heating factor," "Climate," and "Residential Energy Consumption" (blue).

Upon closer inspection of the network map in figure 7, it was discovered that the four clusters were interconnected. For instance, the "NN" term was connected to "Energy consumption prediction" in the same red cluster, as well as "Prediction Model" and "Energy Use" in the yellow cluster. Similarly, it was linked to "Electricity Consumption" and "Smart Meter" in the green cluster, and "Residential Energy Consumption" in the blue cluster. Furthermore, the term "Cluster" in the green cluster was connected to "Prediction Model" in the yellow cluster, "Energy consumption prediction" in the red cluster, and "Residential Energy Consumption" in the blue cluster. Additionally, the terms "Heating" and "Climate" were associated with "Residential Energy Consumption" in the blue cluster, "Cluster" and "Electricity Consumption" in the green cluster, "Prediction Model" in the yellow cluster, and "NN" and "Energy consumption prediction" in the red cluster.

By analyzing the network map in figure 7, the study was able to identify the important terms in each cluster. In the red cluster, the key terminologies included "NN," "Energy consumption prediction," and "SVM". In the yellow cluster, the critical terms were "Prediction Model," "Energy Use," and "Deep NN." In the green cluster, the significant terminologies were "Cluster" and "Electricity Consumption." Finally, in the blue cluster, the essential terms included "Heating," "Climate," and "Residential Energy Consumption."

2.3.3 Analysis of Representative Manuscripts per Topic

To effectively investigate the research inquiries presented, a comprehensive examination of existing scholarly works has been undertaken. Our methodology takes into account several variables that could potentially impact the energy consumption of buildings. In this study, we have conducted an analysis of the energy expenditures of the residents in these structures and have discovered sophisticated computational models capable of effectively categorising or forecasting energy usage. It is worth mentioning that a specific chapter, namely S10, exclusively centres on the identification of electricity consumption trends within buildings. The ACC percentage achieved by this chapter in predicting consumption levels through the utilisation of KMC is 89%. Nevertheless, it is imperative to utilise a model that offers enhanced precision and incorporates supplementary variables that could potentially impact energy consumption. In order to achieve this objective, we propose the development of an intelligent model capable of efficiently forecasting energy use. Both Chapter S11 and S12 focus on the prediction problem, employing different techniques: multiple linear regression and multilayer perceptron, respectively. These models take into account climate conditions and have demonstrated a high level of ACC, reaching rates of 95%, in forecasting electricity and natural gas usage in buildings. It is imperative to utilise models that exhibit a high degree of ACC in predicting energy usage and encompass various influencing elements. As previously mentioned, our chapter examines a curated selection of literature contributions that are in accordance with our research inquiries. During this procedure, a number of IC models that possess the ability to effectively forecast energy usage have been identified and deliberated about.

2.3.3.1 Analysis of metrics, data sources, and critical factors

The research question (RQ1) prompted us to investigate various indicators, data sources, and critical elements that could impact buildings' energy usage. By conducting a comprehensive review of chapters S1 to S41, we were able to identify and extract critical aspects of utmost importance. The investigation of energy usage in residential and public buildings places significant emphasis on concepts such as "electricity," "space heating," and "climate," as depicted in Figure 8. Table 9 displays the primary elements contributing to energy usage in various buildings. The electricity factor was used in 23% of the chapters, while the climate factor was employed in 28% of the chapters. Additionally, space heating was a factor in 23% of the chapters, followed by space cooling in 13%. Lastly, occupant behavior was identified as a component in another 13% of the chapters. By analyzing the components employed as inputs in our research, we identified and observed four distinct points.

The studies (S18, S28, S32, and S41) focused on the electricity factor as the sole determinant in their investigations. For instance, S18 conducted a study to categorize consumer behavior based on the electricity factor, specifically within public buildings. Additionally, S28 proposed a theoretical structure for predicting the hourly energy usage in residential structures. Furthermore, S32 proposed a theoretical framework for predicting energy usage in residential structures using electricity billing data specific to the inhabitants of those structures. In conclusion, the study conducted by S41 introduced a novel approach to categorizing and forecasting energy usage in domestic structures.

Additionally, the study conducted by S8 aimed to decrease energy usage in residential structures. This study exclusively focuses on the climatic element and its influence on energy use. Additionally, it employs a statistical methodology for data analysis to assist decision-makers in conserving energy.

In addition, the studies mentioned earlier (S2, S14, and S22) only focused on customer behavior within residential complexes as the sole variable under investigation.

In addition, S35 utilized the space heating factor to forecast energy usage in public buildings. Ultimately, the remaining portion of the study was predicated upon a combination of variables encompassing electricity, climate, space heating, space cooling, gas, and additional aspects, aiming to categorize and forecast energy usage in public or residential structures.

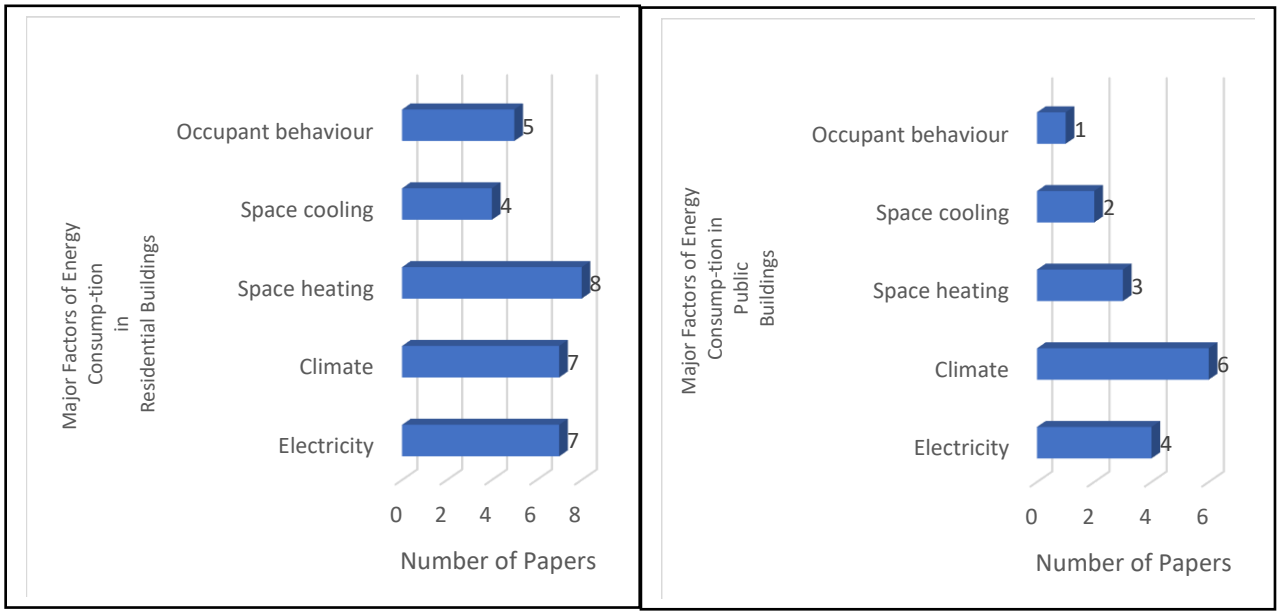


Figure 8. Significant factors of energy consumption in buildings

Table 9. Major factors of energy consumption of buildings

Previous Work	Electricity	Climate	Occupant Behavior	Space Heating	Space Cooling	Gas	Water	Other
S1	-	-	✓	-	-	-	-	✓
S2	-	-	✓	-	-	-	-	-
S3	-	✓	-	✓	✓	-	✓	-
S4	-	-	✓	-	-	-	-	✓
S5	✓	-	-	✓	-	-	✓	-
S6	✓	-	-	-	-	✓	-	✓
S7	-	✓	-	✓	✓	-	-	-
S8	-	✓	-	-	-	-	-	-
S9	-	✓	✓	-	-	-	-	✓
S10	-	-	-	-	-	-	-	✓
S11	✓	✓	-	-	-	-	-	-
S12	-	-	-	-	-	-	-	✓
S13	✓	✓	-	✓	✓	-	-	-
S14	-	-	✓	-	-	-	-	-
S15	-	-	-	-	-	-	-	✓
S16	-	-	-	-	-	-	-	✓
S17	-	✓	-	-	-	-	-	✓
S18	✓	-	-	-	-	-	-	-
S19	-	-	-	✓	-	-	-	✓
S20	-	-	-	-	-	-	-	✓
S21	✓	✓	-	-	-	✓	-	-
S22	-	-	✓	-	-	-	-	-
S23	-	✓	-	-	-	-	-	✓
S24	✓	-	-	-	-	-	-	✓
S25	-	✓	-	-	-	-	-	✓
S26	✓	-	-	-	-	-	-	✓
S27	-	-	-	✓	✓	-	-	✓
S28	✓	-	-	-	-	-	-	-
S29	-	-	-	-	-	-	-	✓
S30	-	-	-	✓	-	-	-	✓
S31	-	✓	-	✓	-	-	-	✓
S32	✓	-	-	-	-	-	-	-

S33	-	✓	-	✓	✓	-	-	-
S34	-	-	-	✓	✓	-	-	-
S35	-	-	-	✓	-	-	-	-
S36	-	-	-	-	-	-	-	✓
S37	-	-	-	-	-	✓	-	✓
S38	-	-	-	-	-	-	-	✓
S39	-	-	-	-	-	-	-	✓
S40	-	✓	-	-	-	-	-	✓
S41	✓	-	-	-	-	-	-	-

2.3.3.2 Analysis of clustering and classification techniques

In addressing Research Question 2, our analysis primarily concentrated on applying clustering and classification methodologies to examine the energy consumption patterns exhibited by buildings. In pursuit of this objective, an examination was conducted on chapters S11 to S20 and S34 to S41. The KMC method is widely utilized in analyzing energy consumption patterns in residential and public buildings, as depicted in Figure 9. Table 10 presents the primary clustering and classification algorithms employed in various buildings. KMC was seen in 76% of the chapters, whereas hierarchical clustering was used in 24% of the chapters. Four key observations were made by examining the clustering approaches employed in numerous research studies. These observations include: Firstly, the studies mentioned earlier (S11 and S13) utilized Ward's method of hierarchical clustering as a means of categorizing energy consumption data in residential settings. Specifically, S11 conducted an investigation to examine the potential decrease in electricity usage and enhancement of energy efficiency within households in the city of Evora, in the southern region of Portugal. The present analysis has successfully identified ten distinct clusters of energy usage. Additionally, a study was conducted by S13 with the aim of investigating the underlying factors influencing thermal comfort behaviors in Portuguese families, particularly concerning cooling and heating practices. This study aimed to establish distinct profiles of power consumption behaviour daily among residential homes. It also categorized families into two fundamental clusters, active and non-active. Furthermore, a study conducted by S12 demonstrated using data mining techniques in smart meters to identify individuals who exhibit greater accountability for peak system usage. This was achieved by analyzing consumption variability and a responsibility factor. In this study, hierarchical clustering and a SOM were utilized to identify the individuals who exhibit greater responsibility inside the peak system of customers.

Secondly, the studies mentioned earlier (S14, S15, S34, S37, and S40) employed KMC with the "KMeans ++" algorithm to categorize the energy usage patterns of buildings. Specifically, S14 focused on investigating the temporal dynamics of household electricity consumption. The electrical consumption was categorized into four distinct groupings. The study conducted by S15 aimed to investigate several aspects influencing power usage, including human activities and the utilization of air conditioning systems. The patterns of individual electricity usage were categorized into six clusters, with an ACC rate of 89.3%. In their study, S34 proposed a methodology for classifying the energy efficiency of residential buildings in terms of their cooling and heating systems. The cooling and heating energy were partitioned into five clusters with a precision level of 87.8%. The authors of S37 proposed a methodology for categorizing energy usage in residential structures. The classification was performed, categorizing the data into four distinct levels: low, medium, high, and extremely high. The ACC of this classification was determined to be 85.9%. The study conducted by S40 aimed to categorize power usage data into five distinct levels: low, medium, high, and very high. The classification ACC achieved by the study was reported to be 90.4%.

Thirdly, as mentioned earlier (S18 and S41), the research employed optimization algorithms alongside KMC technique to effectively categorize electricity usage data within residential structures. The researchers employed optimization methods to ascertain the initial centroid of the KMC

technique. Specifically, in their study, S18 introduced a quadratic programming framework incorporating KMC to classify building energy demand patterns based on occupant behavior. The study's findings revealed ten distinct clusters in power usage, with a precision rate of 83.8%. In their study, researchers at S41 proposed a methodology for categorizing occupant behaviors and quantifying electricity use in various building settings. The results revealed the presence of nine distinct clusters that encompassed different occupant behaviors. The researchers' investigation yielded a classification ACC of 89.7% for inhabitant behaviors in buildings when employing the GA in conjunction with KMC.

Fourthly, the research mentioned above (S16, S17, S19, and S20) utilized KMC and other sophisticated methodologies. In contrast, S16 specifically examined the comparison between two artificial intelligence techniques, namely KMC and hierarchical clustering, to classify energy consumption patterns in residential structures. The present study utilized intelligent meter data to examine the energy consumption patterns exhibited by individuals using a given space. According to the findings of this study, hierarchical clustering demonstrates superior performance in terms of ACC compared to KMC, achieving a rate of 92.8% as opposed to 90.3% for the latter. In their study, the authors of S17 proposed a methodology for categorizing various types of inhabitant behaviors seen in residential structures. This study utilized two primary variables, namely window opening and indoor temperature, and employed two intelligence methodologies, specifically time series analysis and KMC. The study substantiated that KMC outperforms time series in terms of ACC, achieving a rate of 90.20% compared to 87.70% for the latter. In their study, S19 proposed a methodology for categorizing tenant behaviors in residential buildings to ascertain their energy usage patterns. This analysis utilized various parameters, including space heating, refrigeration, and air-conditioning, to determine the projected energy usage. The researchers employed a methodology that integrated KMC with demographic-based probability NNs, identifying ten distinct behaviour consumption patterns shown by residents in residential complexes. The MSE obtained using KMC was 0.09. In their study, S20 introduced a novel methodology for categorizing and examining energy usage in architectural structures, utilizing data obtained from intelligent meter electrical measurements. This study demonstrates the enhancement of KMC performance by using time series analysis and wavelets. The approach that was proposed successfully identified and classified electricity consumption into 12 distinct groups. Applying KMC resulted in a MSE value of 0.18.

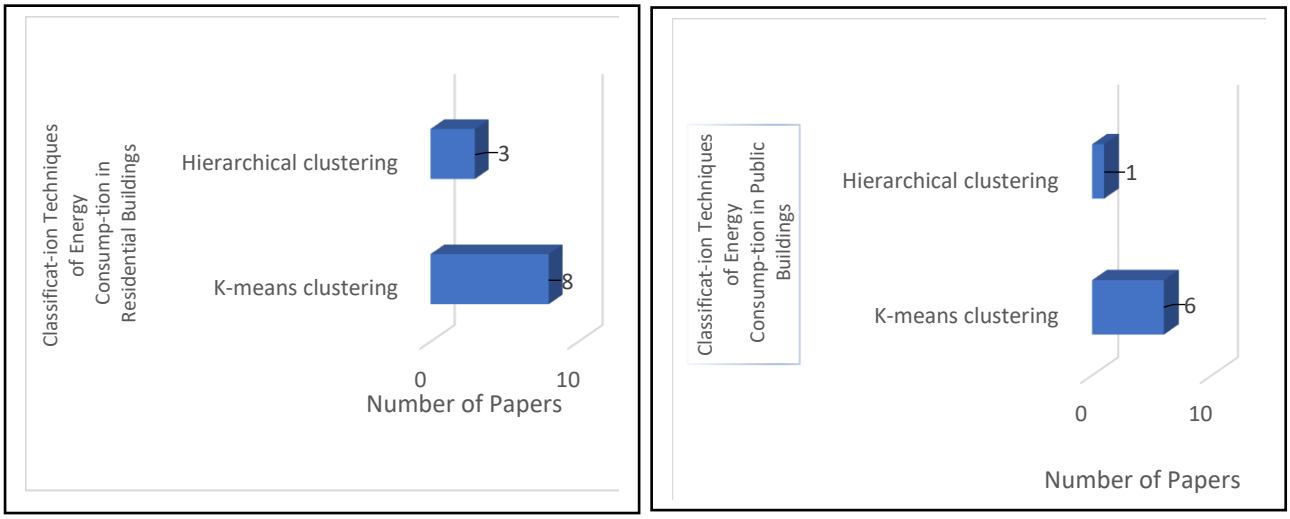


Figure 9. Classification techniques that used energy consumption of buildings

Table 10. Major factors of clustering and classification techniques of energy consumption of buildings

Previous Work	K-Mean Clustering	Hierarchical clustering	Other
S11	-	✓	✓
S12	-	✓	✓
S13	-	✓	✓
S14	✓	-	-
S15	✓	-	-
S16	✓	✓	-
S17	✓	-	✓
S18	✓	-	-
S19	✓	-	✓
S20	✓	-	✓
S34	✓	-	-
S35	✓	-	-
S36	-	-	✓
S37	✓	-	-
S38	✓	-	-
S39	✓	-	-
S40	✓	-	-
S41	✓	-	-

2.3.3.3 Analysis of Prediction Techniques

During the investigation of Research Question 3, we sought out methodologies that may be utilized to forecast the energy consumption of buildings. In pursuit of this objective, an examination was conducted on chapters S21 to S41. The study revealed that the use of several methodologies, including NNs, regression models, and SVM, has been widely embraced, as depicted in Figure 10. Table 11 displays the primary forecasting methodologies employed in various architectural structures. The utilisation of NNs was observed in 35% of the examined chapters, while SVM and regression models were selected in 22% of cases. The analysis revealed that deep NNs were present in 17% of the examined chapters, while the remaining 4% were attributed to the utilisation of the random forest technique. Through an examination of the intelligent prediction methodologies employed in the study, five key observations were made. Firstly, the investigations (S34 and 37) employed the NN technique to forecast energy consumption in residential structures. Specifically, S34 introduced a backpropagation NN approach to predict energy efficiency by considering space heating and cooling, achieving an ACC rate of 85.4%. In addition, S37 proposed a methodology involving the utilisation of a feedforward NN to forecast overall energy usage, achieving a precision rate of 89.2%.

Secondly, the aforementioned research (S23, S28, S41) included NNs and other advanced methodologies. For instance, S23 introduced a predictive model for estimating energy usage in several building categories, including residential, commercial, government, and educational structures. The model employed two machine learning methodologies, specifically ANN and SVM. The NNs achieved an ACC of 90.1%, while the SVM achieved an ACC of 85.4%. According to the authors, NNs exhibit superior ACC compared to SVM. Additionally, a framework was introduced by S28 to forecast the hourly electricity usage in residential structures. The present study utilised sensor data that was gathered from three residential dwellings. Various machine learning techniques, including regression models, feed-forward NNs, and SVM, were employed. The researchers discovered that feed-forward NNs had superior performance compared to the other strategies in terms of MAPE. The MAPE values for the regression, feed-forward NNs, and SVM models were 13.41%, 9.14%, and 9.63%, respectively. In conclusion, the study conducted by S41 introduced a methodology for forecasting occupant behaviours in relation to power usage. This was achieved through the utilisation of a backpropagation

NN and SVM. The algorithms achieved an ACC of 47.02% and 57.14% in forecasting energy load patterns of occupant behaviours, respectively. This demonstrates the superiority of SVM over backpropagation NNs in addressing this particular problem.

Thirdly, the aforementioned research (S21, S26, S29, and S30) employed regression models as a means to forecast energy consumption in diverse building contexts. Specifically, study S21 introduced a methodology for predicting forthcoming energy consumption in a United Kingdom-based supermarket through the utilisation of multiple linear regression. The regression equation is capable of explaining approximately 95.00% of the variation in electricity demand and approximately 86% of the variation in gas use. Furthermore, S26 proposed a theoretical framework for forecasting energy consumption within the context of manufacturing enterprises. The study utilised parameters such as the average outdoor temperature and power statistics. The methodology employed in their study involved the utilisation of multiple linear regression analysis to forecast energy consumption across varying temperature conditions. The findings indicate that the Adjusted R Square value is 0.96. Furthermore, the introduction of S29 implemented a statistical methodology for predicting the overall energy consumption across several sectors, including industrial, commercial, home, and public buildings. The analysis incorporated indicators such as gross domestic product (GDP), population statistics, and GDP per capita. The authors employed a basic regression model as well as multiple linear regression in their analysis. The findings of their analysis indicate that the multiple linear regression model (Adjusted R Square = 0.991) outperforms the basic regression model (Adjusted R Square = 0.844) in terms of the Adjusted R Square metric. In conclusion, S30 conducted a statistical analysis in order to assess the heating energy use of rural buildings. The present study examined various elements, including fundamental family information, characteristics of rural residential buildings, details on the building envelope, interior air quality during winter, and the expenditure of heating energy in buildings. The problem is addressed using a combination of multiple linear regression and logistic regression techniques. The obtained values for Adjusted R Square in logistic regression (0.458) and multiple linear regression (0.471) indicate that multiple linear regression demonstrates superior performance compared to logistic regression.

Fourthly, the aforementioned research (S32 and S40) employed SVM as a means to forecast electricity usage in residential structures, with the suggested model in S32 achieving an ACC rate of 95%, while the model in S40 demonstrated a higher ACC rate of 97.4%.

Fifthly, two studies, namely S24 and S27, employed deep NNs as a means to forecast energy usage in buildings. In their study, S24 introduced a novel framework for forecasting energy consumption in residential buildings. This framework leverages a combination of CNN and LSTM techniques. The MSE for linear regression, LSTM, and CNN-LSTM models were found to be 0.40, 0.74, and 0.37, respectively. Therefore, the CNN-LSTM model demonstrated superior performance compared to both linear regression and LSTM models in terms of MSE. S27 calculated the energy consumption of buildings by considering factors such as space heating and cooling, the energy consumer's classification, and the average consumption of customers. The CNN demonstrated an AER of 31.83 kWh and a SER of 17.29% in its estimations.

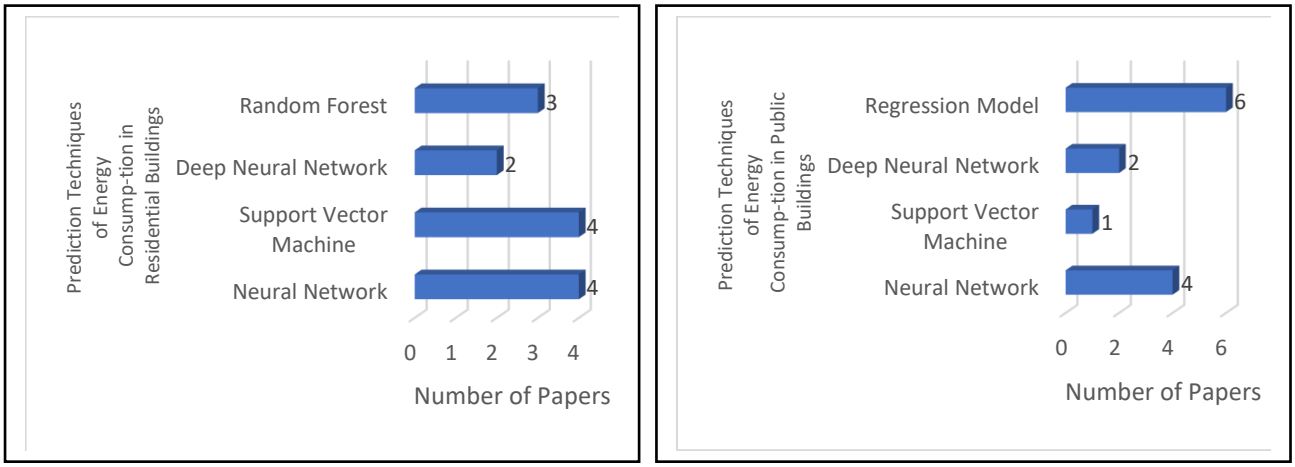


Figure 10. Prediction techniques that used energy consumption in buildings.

Table 11. Significant factors of prediction techniques of energy consumption of buildings

Previous Work	Neural Network	Regression Model	Support Vector Machine	Deep Neural Network	Random Forest	Other
S21	-	✓	-	-	-	-
S22	-	-	-	-	✓	✓
S23	✓	-	✓	-	-	-
S24	-	-	-	✓	-	-
S25	-	-	-	-	-	✓
S26	-	✓	-	-	-	-
S27	-	-	-	✓	-	-
S28	✓	✓	✓	-	-	-
S29	-	✓	-	-	-	-
S30	-	✓	-	-	-	-
S31	-	-	-	✓	✓	✓
S32	-	-	✓	-	-	-
S33	-	✓	-	-	✓	✓
S34	✓	-	-	-	-	-
S35	✓	-	-	-	-	-
S36	✓	-	-	-	-	-
S37	✓	-	-	-	-	-
S38	✓	-	-	-	-	-
S39	-	-	-	✓	-	-
S40	-	-	✓	-	-	-
S41	✓	-	✓	-	-	-

2.3.3.4 Analysis of techniques combining classification and prediction

To effectively investigate Research Question 4, our study centred on the identification of combination methodologies that had the ability to precisely forecast and categorise energy usage in architectural structures. The analysis focused on Chapters S34 to S41 in order to ascertain the most prominent combination approach. The findings of our investigation indicate that the coupling of NNs with KMC is extensively utilised, as depicted in Figure 11. Table 12 illustrates the main combinations

of prediction and classification approaches that were employed in various buildings. Upon doing an analysis of the intelligent prediction and classification approaches employed in the aforementioned investigations, three salient observations were made. The investigations conducted by researchers S35 and S38 employed a hybrid approach that combined KMC with NNs in order to estimate energy usage across different buildings. The authors of S35 proposed a methodology for forecasting heating energy usage. The proposed methodology integrates radial basis NNs and KMC techniques in order to accurately assess the energy efficiency of buildings. The KMC algorithm is utilised to create distinct subsets, which are then used to train unique radial basis function NNs. This approach aims to enhance the ACC of predictions. In their study, S38 proposed a methodology aimed at enhancing the energy efficiency of public buildings in Croatia. In addition, this chapter introduces the use of KMC and a backpropagation NN as approaches to address the aforementioned challenge. The purpose of the study was to investigate the potential impact of KMC on the predictive ACC of a backpropagation NN. Nevertheless, the findings indicate that the integration of KMC and backpropagation did not yield a significant improvement in the predictive ACC of the methodology. Specifically, the utilisation of backpropagation in isolation achieved a 90.1% ACC rate, but the amalgamation of both techniques resulted in a marginal increase to 90.4% ACC.

This study offered two chapters, specifically S39 and S36, to investigate the application of artificial intelligence (AI) methods in the classification and prediction of energy consumption in residential structures. In their study, S39 introduced an innovative methodology that leverages socio-economic variables to forecast energy consumption patterns in residential areas. The application of KMC was utilised for the purpose of analysing load patterns, while the entropy-based feature selection method was employed to find the socio-economic aspects that have an impact on customers' energy load patterns. In addition, the chapter employed a deep NN to forecast the residential load pattern, and the proposed methodology yielded a MSE of 0.12. The study conducted by S36 proposed a hybrid methodology for the classification and prediction of energy consumption in residential structures. This approach combines two artificial intelligence techniques, specifically backpropagation NNs and decision trees. The classification of energy consumption levels was accomplished through the utilisation of a decision tree, while the prediction of energy consumption in residential structures was performed using a backpropagation NN. The decision tree exhibits an ACC rate of 83.6%, whereas the backpropagation NN achieves a higher ACC rate of 91.2%. In conclusion, the second and third questions were addressed by examining studies S34, S37, S40, and S41.

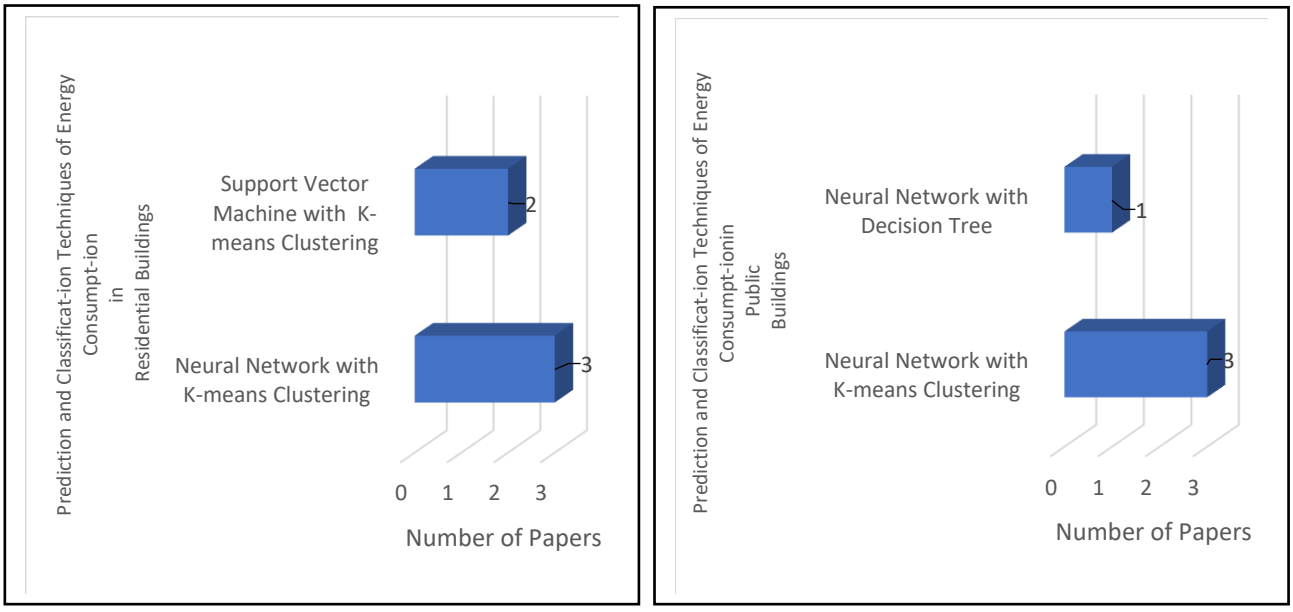


Figure 11. Prediction and classification techniques that used energy consumption in buildings.

Table 12. Major combination prediction and classification techniques of energy consumption of buildings

Previous Work	Neural Network with K-Mean Clustering	Support Vector Machine with K-Mean Clustering	Neural Network with Decision Tree	Other
S34	✓	-	-	-
S35	✓	-	-	-
S36	-	-	✓	-
S37	✓	-	-	-
S38	✓	-	-	-
S39	✓	-	-	✓
S40	-	✓	-	-
S41	✓	✓	-	-

2.3.3.5 Analysis of Performance Evaluation Metrics

Our literature review investigated a range of performance evaluation metrics within the context of our fifth research question (RQ5), as presented in Figure 12. ACC&PRE&REC was employed by 12% of the selected chapters (specifically, chapters S34, S35, S36, S37, and S38). In contrast, 22% of the chapters (S15, S16, S17, S18, S21, S22, S32, and S40-S41) exclusively adopted ACC (ACC). MSE was utilized by 15% of the analyzed chapters (S19, S20, S24, S25, S28, and S39), while only one chapter (S7) employed both ACC and MSE. Adjusted R Square measure was employed by 10% of the chapters (S26, S29, S30, and S33). Interestingly, 39% (16 chapters: S1 to S14, S23, and S31) did not employ any defined evaluation metric.

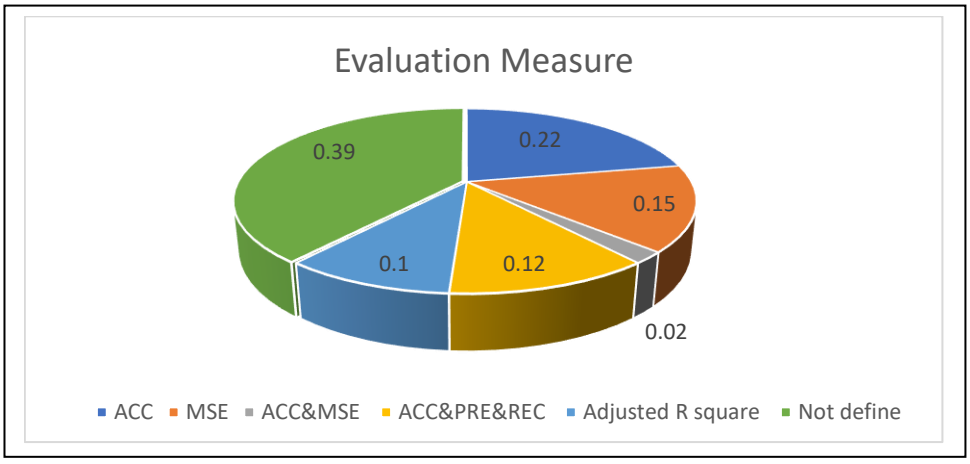


Figure 12. Evaluation Measure of IC Models

2.4 Discussion

This part provides an overview of two significant subjects: our research inquiries and the identification of certain research deficiencies.

2.4.1 Research Question Discussion

The objective of our systematic review was to address five fundamental inquiries pertaining to the use of intelligent strategies in energy consumption across various building sectors.

With regards to our research question 1, we are able to emphasise the primary findings derived from our comprehensive study. The concepts of "electricity," "heating," and "climate" appear to be the most pertinent factors to consider when examining the energy usage of buildings. A parallel was also noted between these findings, derived from a comprehensive examination of 41 chapters, and the outcomes of text mining, encompassing an investigation of 106 chapters. In the text mining methodology employed, TITs were computed for the phrases "electricity" and "heating," yielding values of 0.1 and 0.07, respectively. These results indicate a significant level of relevance for both terms. However, there exists a discernible disparity in the examination of the concept of "climate." Based on a comprehensive examination of 41 chapters, our study indicates that the concept of "climate" holds significance. Based on the findings obtained by text mining, it is evident that the TITs value associated with the term "climate" is 0.02. This outcome indicates that the significance of the "climate factor" is somewhat diminished. In conclusion, it was noted that various additional factors, as depicted in figure 6 as "others," such as socio-economic conditions, geospatial aspects, building characteristics, electricity usage, heating and cooling systems, occupant behaviour, gas consumption, and climate conditions, play a significant role in addressing the issue of energy consumption in buildings. This finding is illustrated in figure 13. In the bibliometric map, a significant correlation was detected between the variables "Residential Energy Consumption" and "Electricity Consumption," "Heating," and "Climate" (refer to Figure 7).

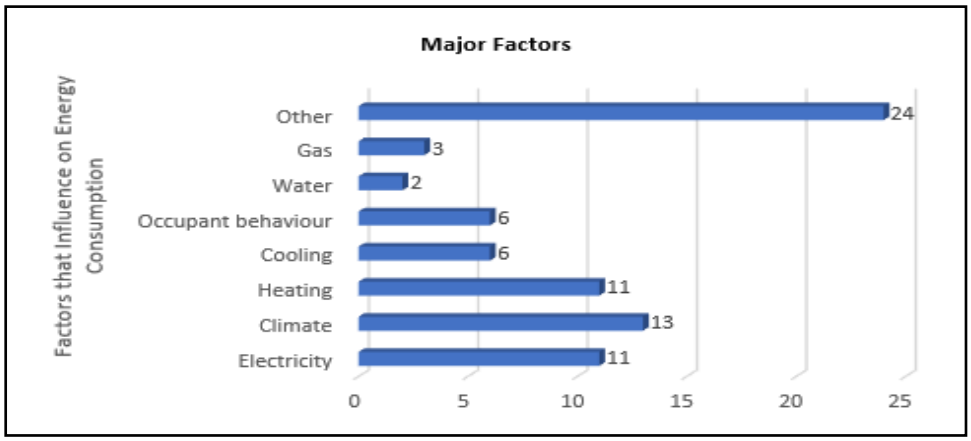


Figure 13. The most relevant factors that influence the energy consumption of buildings from our survey

In relation to Research Question 2, a comprehensive analysis of chapters S11 to S20 and S34 to S41 reveals that the concept of "cluster" holds significant relevance in the context of employing machine learning methodology studying energy consumption patterns in buildings. A correlation was identified between the aforementioned outcome and the inferences derived from the process of text mining. The TITs for the term "cluster" were found to be 0.1, indicating a high level of relevance for this term. In the graphical representation depicted in Figure 6, the concept of "cluster" is subdivided into two distinct words, namely "KMC" and "hierarchical clustering." Upon examination of the aforementioned chapters, it is evident that the utilisation of "KMC" surpassed that of "hierarchical clustering" by a factor of 3.5, as visually depicted in figure 14. The bibliometric map reveals a notable correlation between the "Cluster" concept and the domains of "Electricity Consumption," "Energy consumption prediction," and "Residential Energy Consumption" (refer to Figure 7).

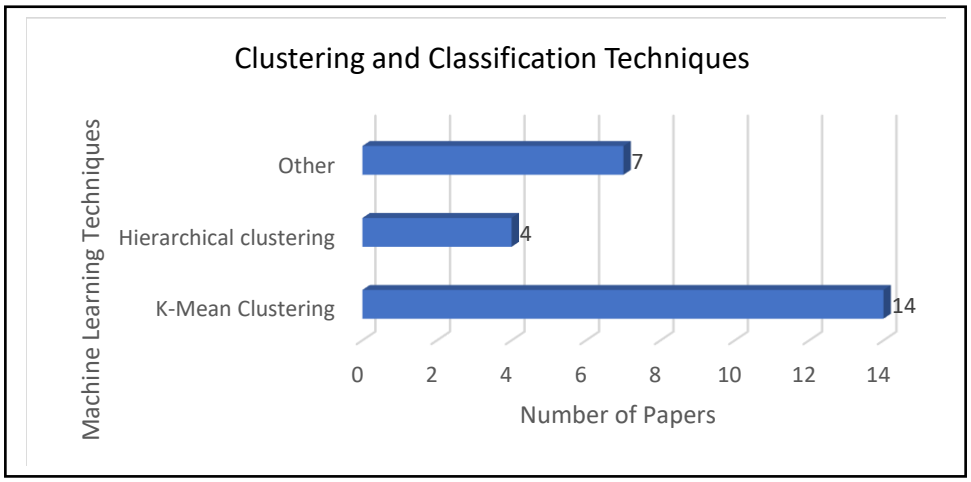


Figure 14. Top classification techniques identified in our survey.

In relation to Research Question 3, our examination of chapters S21 to S41 revealed that the concepts of "backpropagation," "feedforward NN," and regression models such as "multiple linear regression" and "SVM" emerged as the more pertinent terminology. Similar to our examination of Research Question 2, we have identified a parallel between these findings and the inferences drawn from the practise of text mining. The TITs for the concepts "neural" and "regression" were found to be 0.06 and 0.05, respectively. Hence, the phrases "neural" and "regression" are also pertinent. In the range of chapters S21 to S41, the utilisation of NN methodology exceeded that of regression approaches by a factor of 2.5 and surpassed SVM by a factor of 2.143, as visually depicted in figure 15. The bibliometric analysis revealed a notable correlation between the terms "NN" and "Deep NN" with respect to the concepts of "Energy use," "Energy consumption prediction," and "Prediction Model." Furthermore, a correlation may be observed between the "SVM" algorithm and the forecast of energy usage, as depicted in Figure 7.

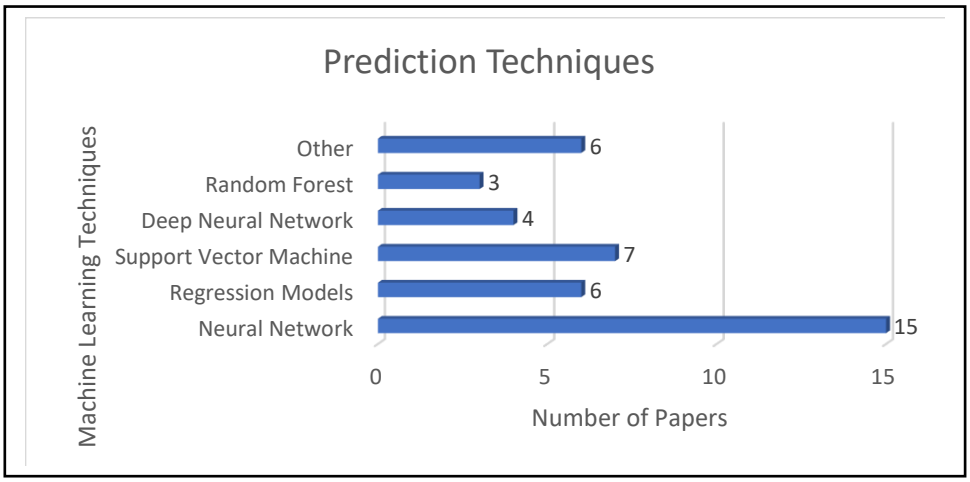


Figure 15. Top prediction techniques identified in our survey.

About Research Question 4 (RQ4), our examination of chapters S34 to S41 revealed that the use of **KMC** was observed in conjunction with backpropagation and feedforward **NN**s in 75% of the aforementioned chapters (namely, in chapters S34, S35, S37, S38, S39, and S41).

About our research question 5 (RQ5), it is crucial to note that the ACC scale serves as a significant metric employed for assessing the precision of the intelligent model employed in analysing the energy consumption patterns across various buildings (refer to figure 12).

2.4.2 Research Gap Discussion

Upon conducting a comprehensive analysis of the aforementioned studies, our survey has identified three primary issues that were addressed.

The absence of a formal investigation hinders the identification of key determinants of energy use across various construction sectors. Furthermore, the manner in which these parameters interconnect with various applications within the realm of energy remains ambiguous. Several studies (S18, S28, S32, and S41) exclusively focused on the electricity factor when examining energy use in buildings. Furthermore, the papers (S2, S14, and S22) only focused on customer behaviour in residential complexes as the sole determinant under investigation. The findings of our study indicate a strong correlation between heating and climate factors and the energy consumption of residential buildings. Similarly, the ECPB is directly influenced by electricity and climate elements.

Moreover, a majority of prior research (S11, S13, S14, S15, S34, S37, and S40) employed KMC and hierarchical clustering methodologies for the purpose of categorising the energy consumption patterns exhibited by buildings. Our investigation reveals that past research has overlooked two key aspects: a) The relationship between "KMC" and "electricity consumption" in public and residential buildings is established. Our chapter suggests the inclusion of alternative classification methods, such as a SOM, to facilitate a comparative analysis with other classification models identified in our survey, with a focus on ACC.

Furthermore, a majority of the preceding research endeavours have employed four fundamental intelligent computer models for the purpose of forecasting the energy consumption of buildings. The models utilised in this study encompass many methodologies, such as NNs (S23, S28, S34, S37, and S41), regression (S21, S26, S29, and S30), SVM (S32 and S40), and deep learning (S24 and S27). Our research demonstrates a clear correlation between the utilisation of NNs and SVM in the accurate prediction of energy consumption in residential structures. Furthermore, a clear correlation exists between the field of "deep learning" and its application in accurately forecasting energy usage in public buildings. Our research suggests employing contemporary methodologies discovered in recent literature, such as recurrent NNs, to contrast and evaluate the prediction models identified in our survey.

Finally, our investigation enabled us to discern various lacunae in the existing body of research. It was determined that a limited number of chapters (S1 and S6) are dedicated to the examination of distinct elements that impact the energy consumption of buildings. Several studies (S5, S11, S18, S21, S26, S28, S32, and S41) examined the electricity factor while neglecting to indicate the occupancy levels or the specific activities conducted within the buildings. A limited number of studies (S31, S33, and S38) have been conducted on the topic of energy usage in public buildings. Nevertheless, stakeholders involved in public buildings, namely in Portugal, perceive this subject matter as significant. They not only demonstrate a willingness to enhance the energy efficiency of these facilities but also exhibit an inclination to switch energy suppliers when market conditions are favourable for such a transition. The findings of this systematic review also revealed a significant knowledge gap in the field of IC models, specifically in relation to the automated classification and prediction of energy consumption in buildings. This gap is primarily due to the extensive range of machine learning techniques currently available in the state-of-the-art literature. Based on the results of our survey, it has been determined that certain strategies with high potential are now underutilised. There exists a limited number of research (S24, S27, S31, and S39) that specifically investigate the use of the deep

NN model, a highly promising methodology for the accurate prediction of energy consumption in buildings.

2.5 Limitations in our Literature Review

There are various limitations inherent in our poll. Significantly, the limitations of this study were mostly attributed to the selection of search keywords and the temporal scope of the articles, which encompassed the most recent five-year period. Furthermore, the study utilised a limited set of electronic databases as sources of information. Moreover, it should be noted that this particular chapter only focused on the analysis of English manuscripts. It is important to acknowledge that there may exist additional available and valuable material that was not included in our examination.

2.6 Conclusions and Future Work

This chapter presents a comprehensive literature analysis that employs a systematic approach to identify and predict the energy consumption of buildings. The primary objective of this review is to address five specific research topics. The application of text mining techniques was employed to identify the frequently utilised phrases within the domains of energy and IC models. Subsequently, a bibliometric map was employed to discern the interrelationships among the most prevalent terms within these domains, preceding a comprehensive examination of the manuscript. Following the PRISMA methodology, our study commenced by identifying a total of 822 manuscripts, which were subsequently subjected to analysis, resulting in the examination of 41 manuscripts. The survey conducted revealed that the predominant IC models employed by the community to classify and predict energy usage in buildings are primarily machine learning methods. This study offers contributions in three distinct areas. The initial study examines the various elements that impact the energy usage of buildings. The second source offers a comprehensive examination of the classification and prediction strategies methodically employed within that particular context. The final factor pertains to the assessment criteria employed by these procedures.

As previously indicated, the study has not encompassed the entirety of publications published in 2021, potentially excluding newly developed intelligent models. The potential enhancement of categorization models and energy consumption prediction in various building sectors can be facilitated by the introduction of novel intelligent techniques.

Therefore, there are still opportunities for improvements regarding our research subject. As a recommendation for future scholarly investigation, our survey offers the exploration of the classification of buildings' energy use by including clustering and optimisation approaches. The primary aim is to proficiently classify the energy consumption patterns exhibited by buildings into discrete tiers, specifically denoted as low, medium, and high. This work suggests the utilisation of machine learning methodologies developed from deep learning techniques for the purpose of predicting energy usage in buildings. The use of specific methodologies such as LSTM, CNN, and deep forest is highly encouraged due to their prominence and current prevalence in scholarly investigations.

Chapter 3 - A Proposed Intelligent Model with Optimization Algorithm for Clustering Energy Consumption in Public Buildings

A. Abdelaziz, V. Santos and M. S. Dias, "A Proposed Intelligent Model with Optimization Algorithm for Clustering Energy Consumption in Public Buildings". International Journal of Advanced Computer Science and Applications, 14(9), 136-152. [15]. https://doi.org/10.14569/IJACSA.2023.0140915.

In recent times, there has been a notable increase in the utilisation of intelligent applications for energy management in public buildings. This can be attributed to their capacity to improve the performance of energy consumption. The management of energy in these buildings is a significant problem as a result of their unpredictable energy consumption patterns and the lack of established design principles for enhancing energy efficiency and promoting sustainable solutions. Hence, it becomes imperative to conduct an investigation of energy consumption patterns in public buildings. This underscores the importance of comprehending and categorising energy usage patterns within these structures. The objective of this study is to identify the most effective intelligent technique for categorising energy consumption in public buildings into distinct levels, namely low, medium, and high. Additionally, the study aims to identify the key factors that significantly impact energy consumption. Lastly, the study aims to establish scientific rules, specifically If-Then rules, that can assist decision-makers in determining the energy consumption level for each building. In order to accomplish the aims of this research, the utilisation of correlation coefficient analysis was employed to ascertain the pivotal factors that impact the ECPB. Additionally, two intelligent models, namely SOM and Batch-SOM based on PCA, were utilised to determine the quantity of energy consumption patterns clusters. The SOM has superior performance compared to the Batch-SOM concerning quantization in ACC. The q-error values for SOM and Batch-SOM are 8.97 and 9.24, respectively. The application of GA was combined with the KM method to forecast cluster levels within each building. The extraction of If-Then rules has been conducted through the analysis of cluster levels. Subsequently, decision-makers are required to identify the buildings that consume the highest amount of energy. Furthermore, this research aids decision-makers within the energy sector in rationalising the energy consumption patterns of inhabitants in public buildings during peak energy consumption periods, as well as facilitating the transition to alternative energy providers for those structures.

3.1 Introduction

The construction industry is currently facing challenges in meeting the rising energy demands, despite its endeavours to promote the development of sustainable buildings (Nguyen & Aiello, 2013). Hence, it is imperative to enhance energy efficiency and conduct a thorough examination of energy consumption trends in buildings. This highlights the need of comprehending and categorising energy use trends within buildings. According to Zhang and Bai (2018), the improvement of building energy quality evaluation is directly proportional to the ACC and practicality of computed energy consumption profiles. According to Javaid et al. (2017), energy usage is influenced by various factors such as building characteristics, costs, and climate conditions. Hence, the task of categorising the energy usage of buildings necessitates the utilisation of sophisticated computational intelligent methodologies. Specifically, the incorporation of cutting-edge machine learning techniques, such as deep learning, proves advantageous as it leverages insights derived from past data. This, in turn, assists decision-makers in the energy sector by establishing a foundation for devising novel strategies for power allocation, particularly in public building areas.

The issue of energy consumption in public buildings is of significant importance, as it constitutes a substantial portion of final energy consumption, as evidenced by data from the ECD countries in 2019. Specifically, within the European Union, this share amounted to 27% (Zhao, Zhong, Zhang, & Su, 2016). As an illustration, it is worth noting that public buildings in Portugal account for approximately one-third of the total electricity consumption, experiencing a notable increase of 35% between the years 1995 and 2019 (Agência para an Energia, 2018). The comprehension of this consumption necessitates the resolution of a multifaceted issue encompassing the physical, technological, and performance attributes of the dwelling, the demographic status, socio-economic factors, climate and weather conditions, and the behaviour exhibited by the occupants of the building (Shi, Liu, & Wei, 2016). Hence, there is a pressing need for assistance in comprehending the energy consumption patterns of public buildings within the context of academic research in European nations, particularly Portugal.

In previous studies, a variety of data mining and machine learning techniques have been employed for the purpose of classifying energy consumption. Clustering is widely recognised as one of the most commonly utilised approaches (Naji et al., 2016). The process of clustering involves the partitioning of objects that exhibit similar characteristics into distinct groups (Massana, Pous, Burgas, Melendez, & Colomer, 2016). Numerous scholarly publications have been produced by researchers pertaining to the categorization of energy use into distinct levels. For example, Gouveia and Seixas (2016) conducted a study in which they utilised a combination of smart metres and door-to-door surveys to identify and analyse household power consumption profiles through the application of clustering techniques. The researcher employed hierarchical clustering methodology to partition household profiles, resulting in the identification of three distinct clusters. In their work, Hernandez et al. (2012) conducted research aimed at classifying daily load curves in industrial parks. To achieve this, they employed a SOM and KM algorithm to calculate the optimal number of clusters. In their study, Ford and Siraj (2013) used a fuzzy c-means clustering algorithm as a technique of categorising smart metre power consumption data into cohesive clusters. In their study, Rhodes et al. (2014) conducted research aimed at categorising residential dwellings based on their hourly electricity use patterns using the KM algorithm. In their study, Azaza and Wallin (2017) proposed a methodology for identifying the most influential energy consumers during peak hours. This approach involved the utilisation of hierarchical clustering and a SOM. In their study, Al-Jarrah, Al-Hammadi, Yoo, and Muhaidat (2017) introduced a technique for identifying the power consumption of buildings by the use of multi-layered clustering. The KM algorithm has been employed for the purpose of partitioning power consumption profiles. Subsequently, the writers ascertain diverse patterns of power consumption profiles. In addition, Cai et al. (2019) proposed a hybrid approach that combines KM with PSO to partition the electricity usage of a given geographical area into many tiers. In their study, Nordahl et al. (2019) employed the centroids of the clusters formed to analyse the patterns of power consumption in households on a daily basis.

The primary focus of scholarly research lies in examining the aggregate energy usage across various buildings through a comprehensive evaluation of relevant literature. Nevertheless, the analysis failed to consider additional variables that influence energy usage, such as the consumption patterns exhibited by inhabitants of these structures at peak periods or periods of vacancy (namely, between 00:00-02:00, 06:00-08:00, and 22:00-00:00). This study aligns with the prevailing literature and presents a novel intelligent computer model for the automated categorization of energy use into several levels, namely low, medium, and high. This model enables the identification of distinct consumption patterns exhibited by public buildings nationwide. These patterns can be visualised at various geographical levels and over the course of a year, highlighting the districts, municipalities, and parishes characterised by low, medium, or high energy consumption during specific time periods. Such insights can assist in guiding occupants' behaviour within these public buildings. Our paper makes a significant contribution in four distinct dimensions:

- The objective of this study is to propose a novel hybrid model, known as the SPKG model, for the classification of energy consumption in buildings, specifically focusing on public buildings. This model integrates various techniques including SOM, PCA, KM, and GA.

- The performance and ACC of the proposed model are assessed through the use of real-world big data pertaining to the ECPB in Portugal. This dataset was gathered over the course of the years 2018 and 2019, encompassing a total of 81,260 public buildings located across 238 locations within Portugal.

- The utilisation of correlation coefficient analysis enables the examination of the association between many elements that impact building energy consumption, with the aim of identifying the most favourable factors for optimal energy efficiency.

- This study aims to develop a clustering and classification model for analysing energy consumption levels in buildings. The model will compare the performance of SOM and Batch-SOM techniques using PCA in terms of q-error. The objective is to identify the ideal model and estimate the optimal number of clusters for analysing energy consumption patterns in buildings. The PCA algorithm is employed to optimise the weights of the SOM, hence improving the fitting capability of the SOM model. Furthermore, the technique of GA was employed to identify the most suitable starting centroids in the KM clustering algorithm. The final technique employed in this study is the prediction of the cluster label assigned to each building.

3.2 Related Work

Categorising public buildings with comparable energy consumption levels is a crucial step in assessing the relative performance of individual buildings within their respective peer groups. Hence, it is crucial to accurately discern these categorizations in order to assist energy decision-makers in addressing three key aspects: optimising energy consumption among high-energy-consuming occupants of public buildings, forecasting future energy demands, and facilitating the transition to alternative energy providers for public buildings.

According to Nordahl et al. (2019), various clustering approaches are frequently employed to analyse energy consumption in buildings. Previous studies have examined unprocessed metre data and employed conventional statistical techniques, including regression analysis and others, to model energy consumption trends (Granell, Axon, & Wallom, 2015; Christ, Braun, Neuffer, & Kempa-Liehr, 2018). The two most commonly employed clustering techniques in the field of occupancy and load forecasting are KM and Hierarchical clustering. These methods have been extensively studied and have demonstrated significant efficacy in this domain (Miller, Nagy, & Schlueter, 2018; Hsu, 2015; Al-Wakeel, Wu, & Jenkins, 2017). Various machine learning techniques, including ANN, SVM, and clustering methods such as K-Shape, have been employed for the prediction of power consumption and loads (Hsu, 2015; Al-Wakeel et al., 2017; Ahmad et al., 2014; Ding et al., 2022; Chen, Xiao, Guo, & Yan, 2023).

Several scholarly research have been conducted to explore the most effective approach for comprehending occupancy schedules and user demand in various buildings. These studies have employed anomaly detection and clustering approaches as their primary analytical techniques (Naji et al., 2016; Hsu, 2015; Cai et al., 2019; Ouf, Gunay, & O'Brien, 2019). Anomaly detection frequently employs occupancy behaviour as a basis for developing tactics that align with dynamic requirements, user circumstances, and interior spatial considerations (Massana et al., 2016). Furthermore, the use of this approach aids in the development of future architectural structures that incorporate a strategic framework for the efficient utilisation of energy resources, as highlighted by Dong et al. (2018).

Several research have been conducted to assess the electricity consumption in buildings and its correlation with various activities. These studies have employed diverse machine learning techniques, such as decision trees (Park, Lee, Kang, Hong, & Jeong, 2016) and stochastic frontier analysis (Aiello, 2018), to measure and analyse the electricity usage. The aforementioned investigations employed sophisticated methodologies to ascertain the various manifestations of electrical loads. Furthermore, this methodology has been used in over 3000 structures, encompassing both residential and non-residential buildings.

Literary endeavours are currently underway to identify an intelligent computational framework for clustering energy consumption in buildings. This framework takes into account various factors that are contingent upon the condition of the buildings at different points in time, with the aim of uncovering patterns in the energy consumption behaviour of occupants within said buildings (Agência para an Energia, 2018). The identification and grouping of energy load patterns exhibited by inhabitants in public buildings can provide significant advantages to stakeholders seeking to enhance the energy efficiency of these structures in a targeted manner. The utilisation of KMC is a prevalent approach within the literature under investigation. However, the aforementioned study highlights various concerns. One limitation of the KMC algorithm is its inability to effectively group data that exhibit variations in volume and density (Park & Son, 2019). Additionally, it has been observed that outliers have the potential to exert influence on centroids (Wen, Zhou, & Yang, 2019). According to Swan and Ugursal (2009), the KM algorithm makes the assumption that all variables possess equal variance. Therefore, our research endeavours to identify a clustering methodology that exhibits enhanced ACC in order to address the inherent constraints of the KMC technique.

Upon careful examination of prior research, it has come to our attention that these studies have encountered difficulties in obtaining comprehensive data pertaining to the behavioural patterns of tenants in various buildings throughout different time periods. Additionally, several research studies employ conventional statistical models such as regression analysis and typical clustering approaches, without adequately examining the efficacy of these models in effectively categorising energy usage into similar groups. The erroneous categorization of energy usage results in various methods that can potentially misguide decision-makers. Firstly, it hampers the ability to identify buildings with high energy consumption. Secondly, it hinders the anticipation of the energy required to adequately meet the needs of public buildings. Lastly, it prevents the identification of the most suitable energy providers. In order to address these limitations, a comprehensive dataset on energy use was gathered from public buildings in Portugal over the years 2018 and 2019. The dataset was employed for the purpose of training and evaluating a hybrid IC model designed to cluster energy consumption patterns in public buildings. This model can serve as a reliable tool for decision-makers in the energy sector to make informed decisions pertaining to energy use in public buildings and energy providers. Furthermore, it is evident that there exists a distinct disparity between the dataset employed in this research and the datasets utilised in prior studies, specifically in relation to data quality and magnitude. This disparity is attributed to the superior quality of comprehensive power consumption data across different time intervals and the substantially larger size of the dataset in comparison to earlier studies. Furthermore, in the preprocessing part, novel hybrid intelligent techniques such as isolation forest and interpolation methods, which have not been employed in earlier studies with the same level of precision and implementation, were utilised. Moreover, a comprehensive analysis has been conducted to identify public buildings that exhibit elevated levels of energy use, surpassing the scope of prior studies. In recent studies, researchers have employed hybrid intelligence methodologies, specifically the combination of Knowledge Management and GA, to effectively forecast cluster labels. The unique characteristics of this study differentiate it from past research endeavours.

3.3 Research Questions and Methodology

In order to establish a solid foundation for our study, we formulated the following research inquiry:

- Research Question 1: What are the many data sources and essential criteria that can be utilised to create a comprehensive profile of the energy consumption patterns exhibited by buildings?
- Research Question 2: Which IC technique(s) may be utilised to ascertain the number of clusters present in the provided ECD?
- Research Question 3: What overarching principles may be derived to assist decision-makers in rationalising energy consumption for public buildings?
- Research Question 4: What are the distinct and fundamental trends identified within the provided dataset on energy consumption?

In order to address the research question at hand, we suggest employing a hybrid methodology that combines machine learning and optimisation techniques. Specifically, we propose utilising the SOM algorithm developed by Massana et al. (2016), along with PCA, KMC algorithm also developed by Massana et al. (2016). This amalgamation of techniques, referred to as the SPKG model, aims to identify diverse energy consumption patterns in buildings. To demonstrate the feasibility of this approach, we present a proof of concept by applying it to public buildings in Portugal. Please refer to Figure 16 for a visual representation of the proposed hybrid approach.

In this section, we describe in detail our proposed model, which is composed of four main phases, as depicted in Figure 16, namely:

- The data collection process encompasses the acquisition of several types of information, including energy consumption data and building features. This includes, but is not limited to, the unique energy point of delivery ID, the address associated with the point of delivery, contractual electrical power, electricity consumption, and billing data with corresponding consumption months. The purpose of this phase is to establish consistency in the units of measurement, provide sufficient sampling rates, maintain the same and synchronised time series, and identify any structural changes that may have occurred throughout the data gathering period.
- The initial stage of the research involves the comprehensive analysis of the data, followed by potential transformations to enhance the information content of the dataset, if necessary. In our study, we utilise various mathematical methodologies, specifically, the Isolation Forest (ISF) methodology for outlier elimination as proposed by Kim, Naganathan, Moon, Chong, and Ariaratnam (2017), and polynomial interpolation as suggested by Aiello (2018).
- Feature Engineering: During this phase, the objective is to identify the most suitable variables for uncovering energy consumption patterns in (public) buildings. This is achieved by employing a coefficients analysis approach, as proposed by Kim et al. (2017).
- In the clustering analysis phase, our study involves the refinement, application, and evaluation of the SPKG hybrid machine learning model. This model is designed to identify clusters, with each cluster representing an energy consumption profile of buildings. Furthermore, the model is utilised to cluster the energy consumption profiles of specific buildings. This study evaluates and contrasts IC methodologies, specifically SOM and KM, in the context of automated cluster identification and the characterization and categorization of energy usage patterns in (public) buildings. Clustering Results: In this phase, we tried to find three significant results: generate energy consumption rules, determine the final number of clusters, and determine municipalities and Portuguese building activities that consume high, medium, and low energy consumption.

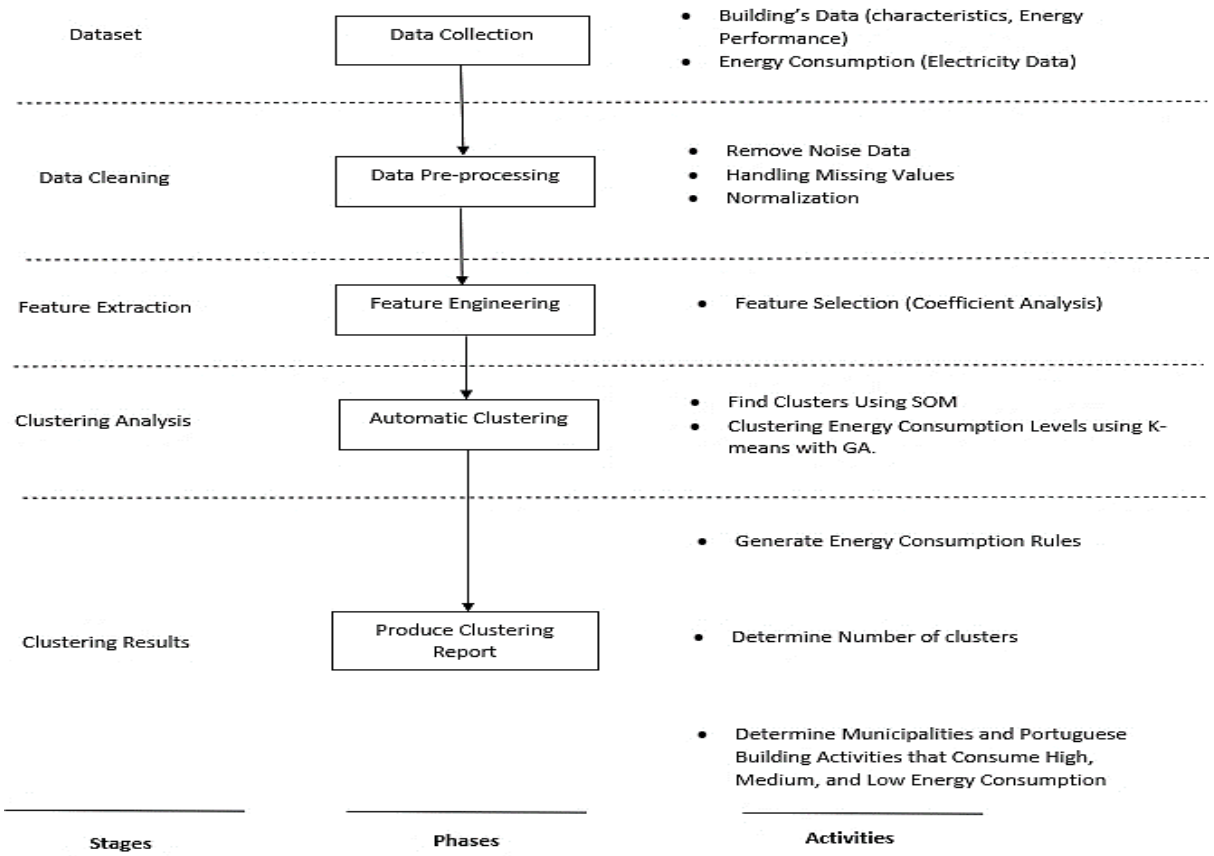


Figure 16. Our Proposed SPKG Model for Discovering Energy Consumption in Public Buildings

3.3.1 Data Collection

This section undertakes an analysis of the energy consumption patterns exhibited by public buildings (NPB) in Portugal. The analysis takes into account the following features: The collection consists of 2,775,082 recordings collected on a monthly basis from 77,996 buildings in various public sectors throughout 238 cities in Portugal, spanning the year from 2018 to 2019. The study utilised a total of 1,222,695 records, which represented 26,624 public buildings. Certain records pertaining to public lighting were excluded from the study due to their irrelevance to the research scope. Additionally, buildings lacking consumption data for the entire 24-month observation period were also excluded. This information is visually presented in Figure 17.

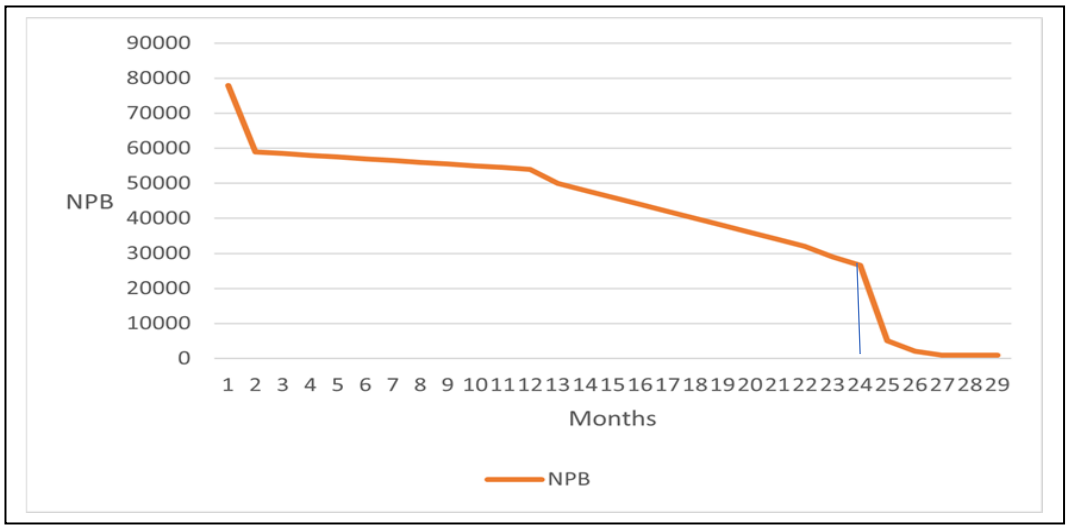


Figure 17. Structure counts in our data collection, with usage months ranging from 1 to 29.

Table 13 presents the traits and dimensions pertaining to the two components comprising our dataset, namely building characteristics and energy use.

Table 13. Dataset Dimensions of Energy Consumption in Public Building

Dataset Dimensions	**Attribute Name**	**Description**
Characteristics of buildings	Unique Energy Point Delivery ID	The ID of each public building
	Business Partner	Identification of the institution that owns or rents the building.
	Building Address	Address of each building
	Municipality	City Location of each building
	Installation Type	Details of the electrical installation of each building
	Contracted Power	Power in MW has been agreed upon with the operator for each building.
	Year/Month	Consumption date
Energy consumption (Active Energy (KWh))	Simple	Total of active energy
	Super Empty	Active Energy (02h00-06h00)
	Empty	Active Energy (00h00-02h00; 06h00-08h00; 22h00-00h00)
	Outside Empty	Lighting and plug loads that cannot be turned off.
	Peak	Active Energy (09h00-10h30; 18h00-20h30)
	Full	Active Energy (08h00-09h00; 10h30-18h00; 20h30-22h00)
	Total	Total of energy consumption (Active plus Reactive Energy)

3.3.2 Data Pre-Processing

This section outlines the methodology employed for data preparation, encompassing the handling of missing data and the elimination of outlier values. The Isolation Forest (ISF) technique is utilised for this purpose. The utilisation of interpolation was implemented in the final portion of this part.

Similar to random forests, the Incremental Supervised Forest (ISF) is built using decision trees. In the absence of externally provided labels, the implementation can be classified as unsupervised. The notion of "limited and distinguishable" data points played a pivotal role in the development of isolation forests for the purpose of identifying anomalies within our dataset. Decision trees were constructed using information criteria such as the Gini index or entropy. Once the prominent discrepancies have been resolved in the lower part of the tree and further up its branches, more nuanced distinctions become apparent. The isolation forest algorithm operates by partitioning the data into a tree structure using randomly determined criteria on randomly subsampled data. There exists a minimal probability that samples located deeper within the hierarchical structure and requiring a greater number of divisions to distinguish them from the rest of the data are outliers. Likewise, it is more probable for samples that are in closer proximity to the root of the tree to exhibit outlier characteristicsthe tree has distinguished these samples tree as distinct from the other data in terms of energy consumption.

ISF consists of two distinct stages. During the modelling phase, random subsets of the ECD are picked to construct the iTrees collection. The evaluation step employs iTrees to conduct tests on data and records the path length for each test instance prior to computing the exceptional outcome. Subsequently, the aforementioned study conducted by de Santis and Costa (2020) focuses on the process of segregating and identifying any anomalous test results.

The construction of decision iTrees occurs during the modelling stage, wherein the given dataset is iteratively divided into segments until all instances are isolated or the tree reaches its predetermined maximum depth (MaxD) (Arjunan, Poolla, & Miller, 2022). The anormality score of each instance is determined by the iTrees acquired during the preceding modelling phase. The following are the details pertaining to this particular stage:

- Let x. go through each iTree in the model, recording its location at the end. At the root of every iTree.
- First, using formulas 3 and 4, calculate the path length of instance x and the abnormality score S; Then, Step 2 uses the abnormality score to estimate instance x (Sternby, Thormarker, & Liljenstam, 2020).

$$S\,(x,n) \;=\; 2\,(\mathrm{E}(\mathrm{h}(\mathrm{x})))\,/(\mathrm{c}(\mathrm{n})) \qquad (3)$$

$$\mathrm{c}(\mathrm{n}) \;=\; 2\mathrm{h}\,(\mathrm{n}-1)-(2(\mathrm{n}-1))\,/\mathrm{n} \qquad (4)$$

The average path length of instance (x) in each iTree is indicated in Eq.(3) by E (h(x)). The average of h (x) is represented by c(n). It is used to normalize h(x). Three situations can be found in E(h(x)) (Hariri, Kind, & Brunner, 2019):

- E(h(x)) equals zero, (s) equals one, which indicates that the likelihood of an abnormality for (x) increses. if (s) score is very close to one.
- This suggests that if (s) reaches 0.5, (x) is not an essential anomaly. E(h(x)) equals c(n), (s) equals 0.
- E(h(x)) equals (n,1), (s) equals zero, which indicates that (x) is more likely to be a standard instance if (s) is smaller than 0.5.

Authors commonly resort to interpolation, a malleable mathematical strategy to compute compensation values based on related known values (Benachour, Draoui, Imine, & Asnoune, 2017). The graphical representation of energy consumption in public buildings can be achieved by extrapolating unknown data points based on a consistent pattern observed within a given dataset.

Upon analysing the findings of the ISF, it was observed that the energy provider incorporated zero and negative values into the recorded consumption data in certain instances. This phenomenon can be attributed to the need for adjusting previous consumption estimates in buildings without smart metres, where frequent manual readings are necessary and may diverge from the real consumption data. Additionally, it was seen that disregarding these values would result in a decrease in the overall count of public buildings to 10361. This reduction in sample size would provide a substantial challenge in the application of our models. In order to address the issue of missing and negative values, it was necessary to employ an interpolation approach for the purpose of imputing consumption values. The Difference Table and the Lagrange method represent two among several interpolation approaches that are readily available. Given the non-uniform spacing between neighbouring data points, we employed the Lagrange method to address the aforementioned issues pertaining to negative and zero values. For example, Building A exhibits a monthly energy consumption of 500 kW in the initial month, followed by 200 kW in the subsequent month, and 280 kW in the third month. Given this premise, it is not appropriate to make direct comparisons between the differences seen in the three months.

The Lagrange technique was employed in order to accommodate negative and zero values inside our dataset. This strategy necessitates the utilisation of three main inputs, as outlined in Algorithm 2.

The various types of polynomial degrees are shown in Table 14 (Benachour et al., 2017). Equation (5) (Abdelaziz, Santos, & Dias, 2021), shows how the RMSE metric was used to determine the polynomial degree. Set n points at $(x_0, y_0), \ldots, (x_n - 1, y_n - 1)$ and provide the corresponding function values in (c) to represent the array of f (a) at each of the n points. Step 6 then involves computing the Lagrange polynomial that is built so that xi is substituted for x to have a value of zero whenever (j ≠ i), and a value of yi when j = i. The Lagrange polynomial obtained by summing these terms has the form $p(x_j) = 0 + 0 + \ldots + y_j + \ldots + 0 = y_j$ for each of the coordinates (x_j, y_j). The interpolation results are then displayed using the equation in step 9 in Algorithm 2.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - o_i)2}{n}} \qquad (5)$$

Where:

- The predicted value (compensation values for numbers less than 1 and 0 values) for the *ith* observation in the dataset is represented by the symbol p_i.
- The observed value (energy consumption dataset) for the *ith* observation in the dataset is denoted by the symbol o_i.
- The size of the sample is n.

Table 14. Polynomial Degree Structure - Various Forms

Polynomials	Degree	Examples
Constant Polynomials	Polynomials of a Certain Degree 0	3
Linear Polynomials	Polynomials of a Certain Degree 1	x + 8
Quadratic Polynomials	Polynomials of a Certain Degree 2	$3x^2 - 4x + 7$
Cubic Polynomials	Polynomials of a Certain Degree 3	$2x^3 + 3x^2 + 4x + 6$
Quartic Polynomials	Polynomials of a Certain Degree 4	$x^4 - 16$
Quintic Polynomials	Polynomials of a Certain Degree 5	$4x^5 + 2x^3 - 20$

The second algorithm involves the utilisation of the Lagrange interpolation method to compute offsets for the negative and zero values inside the ECD.

Algorithm 2: Lagrange Interpolation Method to find compensation values to negative and zero values in the energy consumption dataset

Input: n-Degree, Points:$a_1, a_2, \ldots\ldots, a_n$, Function values: $f(a_1), f(a_2), \ldots\ldots, f(a_n)$, Evaluation point: X

Output: The value of the nth-degree Lagrange interpolant at point X

1. Answer = 0;
2. **For** I to n **do**
3. Product = 1;
4. **For** j to n **do**
5. **If** I ≠ j
6. Product = (Product) × $\frac{X - a_j}{a_I - a_j}$
7. **End**
8. **End**
9. Answer = Answer + (Product) × $f(a_i)$
10. **End**
11. **Return** Answer

Data preprocessing can be condensed into three primary phases, as outlined below:
The initial step involves the implementation of ISF for the purpose of detecting outliers:

- The task at hand is identifying the specific features (columns) within the dataset that contain missing data.
- To differentiate between the complete data (rows without missing values) and the incomplete data (rows with missing values), it is necessary to examine each attribute individually.
- The isolation forest approach is utilised to isolate outliers by employing the entire dataset for each characteristic. An effective method commonly employed for anomaly detection is known

as isolation forest. This technique involves the construction of random forests and the subsequent determination of the average number of splits necessary to isolate a given data point.
- One should establish a threshold in order to effectively identify outliers. The outcome of this may vary based on the quantity of divisions or a pre-established threshold.

The second step in the data analysis process involves the imputation of missing values:

- Interpolation techniques should be employed in order to infer the missing values pertaining to the features that exhibit incomplete data. Interpolation is a mathematical technique employed to estimate the values that are absent within a given dataset by utilising the existing data points.
- The user's text does not contain any information to rewrite. There exist various interpolation techniques, such as linear interpolation, polynomial interpolation, and specialised methods designed for time series data, such as forward-fill and backward-fill.

The next step in the process involves the integration of outlier detection and imputation techniques:

- After the identification process using the isolation forest, we made the decision to retain the outliers within the dataset.
- The chosen method of polynomial interpolation will be utilised to complete the data gaps by estimating the missing values.

3.3.3 Feature Selection

This section aims to find the critical variables or factors in our energy consumption dataset. To overcome this problem, we used the T-test correlation coefficient. This statistical technique is used in literature to detect if 2 factors/variables are significant (H. X. Zhao & Magoulès, 2012). It can be helpful in our study. In our dataset, looking at pairwise correlations between the various variables (or factors) may propose a causal relation between 2 factors that we can investigate further. Equation 6 computes the T-test value by assuming no correlation with ρ = 0. Where P refers to that, there is no relationship between variables (Müller, 2021).

$$t = r\sqrt{\frac{n-2}{1-r^2}} \qquad (6)$$

In (6), n refers to the instances, and r represents the correlation coefficient of the energy consumption dataset. The importance of relevance is expressed in probability levels: p (e.g., significant at p = 0.05). The degree of freedom for entering the t-distribution is n – 2. If the t value is less than the critical value (CV) at a 0.05 significant level, the factor is not essential and is avoided (Müller, 2021).

In Algorithm 3, the correlation coefficients are constructed using the training dataset. The correlation coefficients between the proposed components are calculated in Steps 1 to 4. The computation of significant values from step 6 to step 7 is achieved by the utilisation of the T-test. Ultimately, steps 8 through 10 culminate in the determination of the definitive compilation of energy consumption factors.

```
Algorithm 3: Feature Selection Algorithm
Input: S(F_1, F_2, ……., F_k, F_c) // a training data set
Output: S_best // the selected feature set
1. begin.
2. For I to k do
3.      r = compute correlation coefficient (F_i,F_c)
4. End
 // let P = 0.05 significant level
let P = 0 // assuming there is no significant correlation
5. For I to k do
6.       t = compute significant values (r,p) for Fi  // Eq.4
7.       If t > CV // critical value
8.          S_list = CV
9.        S_best = S_list
10.     End
11.End
12. Return S_best
```

3.3.4 Finding the Number of Clusters

In order to ascertain the most suitable number of clusters in ECD, three established approaches from the literature were employed: the SOM, the Elbow method, and the Bouldin & Davis method (Azaza & Wallin, 2017; Al-Jarrah et al., 2017). Previous research has employed these methodologies to determine the most suitable number of clusters, particularly in the context of building energy usage.

3.3.4.1 Self-Organizing Map

The SOM is a type of NN that is widely used as a tool for clustering and visualising data in the field of exploratory data analysis (Lee, Kim, & Ko, 2019). The primary goal of SOM is to transform a complex input space with high dimensions into a lower-dimensional output space while preserving the topological relationships among the data points, but not the actual distances between them (Ioannou, Kofinas, Spyropoulou, & Laspidou, 2017; Liu & Ren, 2018). The utilisation of an unsupervised learning algorithm involves the implementation of a fundamental heuristic approach to identify concealed non-linear patterns inside datasets characterised by a large number of dimensions (Ioannou et al., 2017).

The adoption of the SOM approach is justified due to its accurate and effective handling of large datasets (Lee et al., 2019). In contrast to alternative approaches, this strategy demonstrates enhanced efficacy in handling datasets of smaller to medium sizes (Ioannou et al., 2017). Hence, the SOM technique was employed to ascertain the optimal number of clusters within the ECD. It encompasses three primary processes, including competition, collaboration, and adaptation, as outlined by Lee et al. (2019).

The SOM network comprises two levels, namely the input layer and the output layer, as seen in Figure 18 (Lee et al., 2019). The m-dimensional input vector is used to represent each input variable (Abdelaziz et al., 2021). The number of nodes in the output layer of a SOM has a significant influence on the precision and generalisation capability of the SOM, as highlighted by Lee et al. (2019). Additionally, the number of nodes in the output layer corresponds to the maximum number of clusters that may be formed by the SOM. The process of initialising the weight vectors is the initial step in the organisation of the SOM (Ioannou et al., 2017). then, the weights are interconnected to facilitate the connection between the input nodes and the output nodes and are then adjusted through the process

of learning. In order to determine the BMU, the distances between an input (x) and the weight vectors (wi) of the SOM are computed using several measurement techniques, as described in previous studies (Liu & Ren, 2018; Räsänen et al., 2008).

- MD.
- CD.
- ED.
- Mahala Nobis distance
- Vector product, among other methods.

Euclidean distance is an approved measure in most scientific papers (Liu & Ren, 2018), as shown in Eq. (7):

$$(t) = \|x\ (t) - \text{wi}\ (t)\ \| \text{ di} \qquad (7)$$

At the finish of the propinquity matching method (Determine the similarity between points in the dataset), the most excellent matching unit c at repetition t is identified by the minimum distance (Lee et al., 2019).

$$c\ (t) = \arg\min i\ \|\text{di}\ (t)\ \| \qquad (8)$$

By analyzing the weight vector wi(t) of the winning neuron, i at iteration t, the overhauled weight vector wi(t + 1) at iteration (t + 1) is determined by Using a discrete-time formalism formalism in in Eq (9) (Liu & Ren, 2018).

$$\text{wi}(t+1) = \text{wi}(t) + \alpha(t)\ [x\ (t) - \text{wi}\ (t)] \qquad (9)$$

The weights (α) adjustment rate diminishes away from the winning node regarding the Spatio-temporal decay function (Ioannou et al., 2017).

$$h_{ci}\ (t) = \exp\ (-(d*d)^{ci} / 2\sigma^2\ (t)) \qquad (10)$$

where:

- d is the lateral distance between the winning neuron c and the excited neuron i.
- sigma is the effective width or radius of the neighbourhood at iteration i.

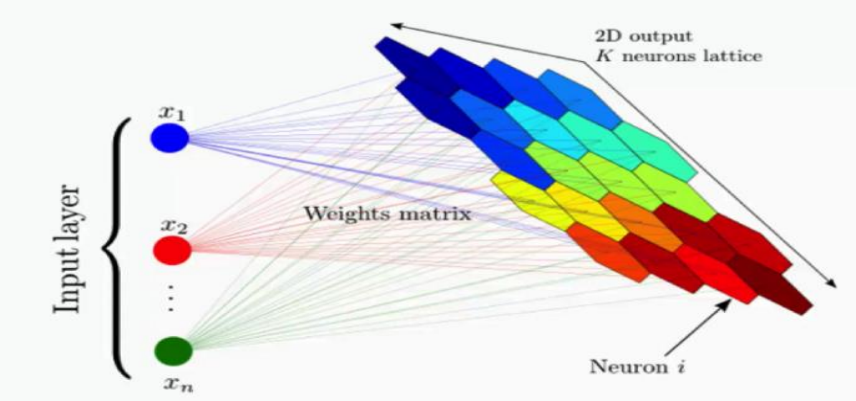


Figure 18. Structure of SOM

Algorithm 4 was utilised to create a SOM network using energy usage data in order to ascertain the most favourable number of clusters. To begin, the lattice space of a 10 × 10 grid was determined. Random weights and PCA weights were assigned to the lattice. The number of iterations was varied from 100 to 1000. Furthermore, the next step involves selecting random points from the energy consumption data. These points will be used to identify the best match point based on Equation (10). The learning rate is set to 0.5, and the neighborhood function is defined as a triangle. Subsequently, the neighbourhood distance weight matrix is computed and used to modify the SOM weight matrix. This process is then repeated, starting from the step of selecting a random point (z), until the maximum number of iterations is reached.

Algorithm 4: Main Idea of the SOM Network Training
Input: ECD ← the energy consumption data. **Output:** USOM ← U-matrix of SOM network. 1. β← initialize lattice nodes. 2. Ω← initialize weight vectors. 3. N ← Iteration count. 4. **For** i ← 1 to N do 5. z ← picks a random point in ECD. 6. c ← β closest to z. 7. move the weight vector of c closer to z. 8. move the weight vectors of the neighbours of c slightly closer to z. 9. **End** **Return** USOM

3.3.4.2 Elbow Method and Bouldin-Davis Method

We can plot the curve indicating the average inner per cluster sum of squared error (SSE) distance vs the number of clusters to discover a visual "elbow", the ideal number of clusters. The average inner whole of squares is the average distance between focuses interior of a cluster (Hernández et al., 2012), as shown in Eq. (11)

$$W_k = \sum_{r=1}^{k} \left(\frac{1}{n_r} + D_r \right) \qquad (11)$$

Where:

- k is the number of clusters,
- n_r is the number of points in cluster r.
- D_r is the sum of distances between all points in a cluster.

In Davis and Bouldin (DB), the score is characterized as the average similitude degree of each cluster with its most identical cluster. The similitude is the proportion of within-cluster separations to between-cluster separations. In this way, clusters that are more distant separated, and less scattered will result in a distant better score. The least score is zero, with lower values indicating superior clustering (Rhodes et al., 2014), as shown in Eq. (12) and (13) (Cai et al., 2019).

$$DB(c) = \frac{1}{k} \sum_{i=1}^{k} \left(max_{j \le k, j \neq i} D_{ij} \right), k = |c| \qquad (12)$$

D_{ij} is the "within-to-between cluster distance ratio" for the ith and jth clusters.

$$D_{ij} = \frac{di^{-} + dj^{-}}{dij} \quad (13)$$

Where di^{-} is the average distance between every data point in cluster i and its centroid, similar for dj^{-}. d_{ij} is the Euclidean distance between the centroids of the two clusters.

3.3.5 K-Means with GA

Grounded in the conceptual framework established by Charles Darwin's theory of natural evolution, GAs are a computational approach that exhibits versatility in its applicability to various problem domains. Selective breeding is a practise that promotes the reproduction of individuals that possess superior strength and health within a population, hence guaranteeing the birth of offspring that exhibit these desirable traits in subsequent generations. The utilisation of GAs expedites the implementation of various fitness functions, including the ED, within the context of analysing ECDs. This is primarily attributed to the inherent efficiency of GAs in handling large datasets comprising numerous data points, as well as their resilience in effectively addressing noisy conditions. The fitness functions for ED, MD, and CD were utilised in the GA to identify the optimal centroids in KM, hence facilitating the convergence of energy consumption sites. The formulas 14, 15, and 16, as shown by Rhodes et al. (2014), demonstrate the application of these fitness functions. Furthermore, it enhances the ACC of KM's methodology in our approach.

The initial stage of natural selection in GA involves the identification and selection of the most adaptive individuals within a population. Offspring are produced by individuals, inheriting and perpetuating the characteristics of their parents, therefore becoming part of the subsequent generation. Parents that possess a greater level of physical fitness are more likely to give rise to offspring who surpass their own performance and exhibit an increased likelihood of survival. Through numerous iterations, the most physically capable generation will ultimately emerge. A GA encompasses the following five stages: The original population was determined. The method commences by establishing a population group of individuals. Each individual constitutes a vital element in resolving the matter at hand. 2) The concept of a fitness function. The fitness function evaluates an individual's fitness level, which refers to their capacity to compete with others. Each individual is assigned a fitness rating by the system. The probability of an individual being selected for reproduction is determined based on its fitness score. 3) Selection: In the selection phase, the most optimal individuals are identified and have the opportunity to transmit their genetic material to subsequent generations. Two cohorts of individuals (referred to as parents) are selected based on their fitness ratings. Individuals with a high level of physical fitness are more likely to be selected for the purpose of reproduction. 4) Crossover is widely recognised as the pivotal phase in a GA. A crossover point is selected at random within the DNA sequence for each pair of parents to engage in mating. The genetic material of parents is transferred in a process known as crossover until a specific point is reached, leading to the production of children. Mutation is a phenomenon whereby newly generated children have a small probability of experiencing alterations in one or more of their genes. This implies that certain bits inside the bit string have the potential to be reversed. Mutation occurs in order to maintain genetic diversity within a population and prevent premature convergence.

The coordinates (x_1, y_1) of one point are used in the ED formula, while the coordinates (x_2, y_2) of another point are used to calculate the distance between the two points (x_2, y_2).

$$ED = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (14)$$

MD represents the total absolute value of the coordinate differences. Here's an illustration of how to calculate the MD between two data sets: say X = (E, M) and Y = (B, K).

$$MD = |E - B| + |M - K| \qquad (15)$$

CD determines the cosine of the angle formed by vectors X and Y.

$$CD = \frac{X.Y}{\| X \| \| Y \|} \qquad (16)$$

Where:

|| X || = Mean Euclidean Distance of a Vector, X = (X_1, X_2, ..., X_n)

|| X || = An example of a vector's Euclidean norm, Y = (Y_1, Y_2, ..., Y_n)

The methodology employed by algorithm 5 for the computation of optimal KM centroids utilising GA is elucidated in the subsequent scientific exposition. At the initiation of the ensemble, a considerable quantity of chromosomes, referred to as the ECD , are present. The objective of the GA is to identify the optimal ECD centroids by minimising the SER and maximising the chromosomal variance. The computation of ECD's fitness function, which includes the evaluation of ED, MD, and CD, is performed. It is assumed that the execution has concluded following the specified number of iterations. The use of Early Childhood Development (ECD) enables individuals to identify the most appropriate centres for their needs. In this scenario, it is necessary to execute the selection process and select the two most optimal chromosomes (2 ECD) from the population, taking into consideration their fitness function value. Subsequently, a random selection is made from the population, wherein two chromosomes (ECD) are chosen. The subsequent phase involves executing the crossover operation and identifying the exchange point, wherein the parent exchange is a subset of a collection of detachable exchange sites denoted by binary values. Two chromosomes that are not genetically related are randomly chosen from the available pool of probable candidates. Subsequently, the resultant offspring would undergo a mutation operation wherein the bit locations are reversed. In conclusion, an elitist methodology is implemented in order to maintain the persistence of favourable chromosomes (ECD). This involves generating a fresh population from which a fitness function can be derived, and subsequently iterating through the aforementioned processes until the most optimal centroids in KM are identified.

Algorithm 5: Techniques for Locating the Best Centres in KM Using GA:

Algorithm 5: GA Steps for Finding the Optimal Centroids in KM
Input: Size α of population, Number σ of Iterations **Output:** β ← (Optimal Chromosomes $_{(Optimal\ Centroids)}$). 1. Count I = 0 2. C_k = Create random σ solutions 3. Compute fitness function (i) for each $I_\varepsilon C_k$ 4. ***While*** *(σ ≠ 0)* 5. ***For*** *I = 0 to σ do* 6. Pick chromosomes (ECD) for the contest. 7. Detect chromosomes (ECD) with the lowest fitness value. 8. Avoid chromosomes (ECD) with the lowest fitness value.

9. Estimate novel chromosomes (ECD)
10. ***End For***
11. Execute the mutation method.
12. Estimate (mutated chromosomes((ECD)).
13. Compute fitness function (i) for each $I_{\epsilon}C_k$
14. ***End While***
15. β = fitness values from C_k

Return β

One of the fundamental objectives of knowledge management is to cluster similar data points together in order to uncover latent patterns and trends. The subject encounters numerous challenges. One aspect involves determining the optimal number of clusters by striking a balance between the Elbow method and the Davis & Bouldin method. Additionally, the technique of GA is employed to determine the optimal location of the centroid inside each cluster. Due to this rationale, the use of KM has been implemented in order to forecast the labels allocated to clusters including all ECD, so efficiently clustering the extent of energy consumption exhibited by each building. A more advanced KM cluster was constructed by the integration of the SOM, the Elbow methodology, the Davis & Bouldin method, and the GA. In order to enhance the precision of cluster label prediction for all ECD buildings, an enhanced version of the KM algorithm is employed. This upgraded version incorporates stages 1 through 4, which include the identification of new centroid positions inside each cluster. For a detailed description of this process, please refer to method 6.

Algorithm 6: Cluster label predictions in each structure have been improved thanks to enhanced KM:

Algorithm 6: Improved KM to predict cluster labels in each building

Input: K = 3, // Specify the number of clusters using the Elbow method and the Davis & Bouldin method

Initialize σ of centroids using GA.

Output: $\beta \leftarrow$ predicting cluster label in each building in ECD

1. **Repeat**
2. Assign each point to its closest centroid.
3. Compute the new centroid of each cluster.
4. **Until** the centroid positions do not change.

Return β

3.4 Experimental Results and Discussion

This section encompasses four distinct components: data pre-processing, feature selection, determination of the optimal number of clusters, and ultimately, the application of the KM algorithm with a GA to generate guidelines for energy usage. The proposed methods were implemented using Python programming language and the Scikit-learn module.

3.4.1 Results of Data Pre-processing

- The efficacy and use of any study employing intelligent machine learning techniques are contingent upon the soundness and appropriateness of the dataset. Consequently, the ability to design and refine an intelligent model is significantly improved when the provided dataset is of good quality. Furthermore, the data pertaining to energy usage is derived from an authentic real-world context. Consequently, the elimination of noise and outliers is a crucial component of the data preparation phase. Data preparation consists of two distinct processes.

Figure 19 illustrates the preliminary phase of the methodology, involving the preprocessing and exploration of energy use data. Figure (a) displays the relationship between contractual power (Xi) and total energy consumption (Yi) for a specific sample of the raw dataset. Figure (b) illustrates public buildings that have energy consumption periods either shorter or longer than 24 months. The exclusion of these items from our study was based on their deviation from the intended scope of our investigation. In section (c), our dataset exhibits a greater number of negative and zero values. Consequently, we have employed an interpolation approach to replace these values with estimated values in order to maintain the integrity of our dataset. In the fourth scenario (d), the ISF method has been employed to remove outlier values. However, it is important to note that this approach has also resulted in the elimination of potentially hazardous or zero values. In our implementation of the ISF approach, we prioritise the following criteria:

- The ensemble size, or the total number of trees generated, equals the number of estimators (n estimators = 100). There is a default value of 100.
- The number of samples used to train each simple estimator is denoted by the max samples value (max samples = auto). When max samples are set to "auto," it will be set to min (n = 256 samples).
- The contamination (auto) variable in our dataset reflects the expected fraction of extreme values. When the "auto" setting is used, the contamination level is set to 0.1.
- To train a tree, take as many features as possible from the complete set of features, which is represented by Max features (set to 1 by default).

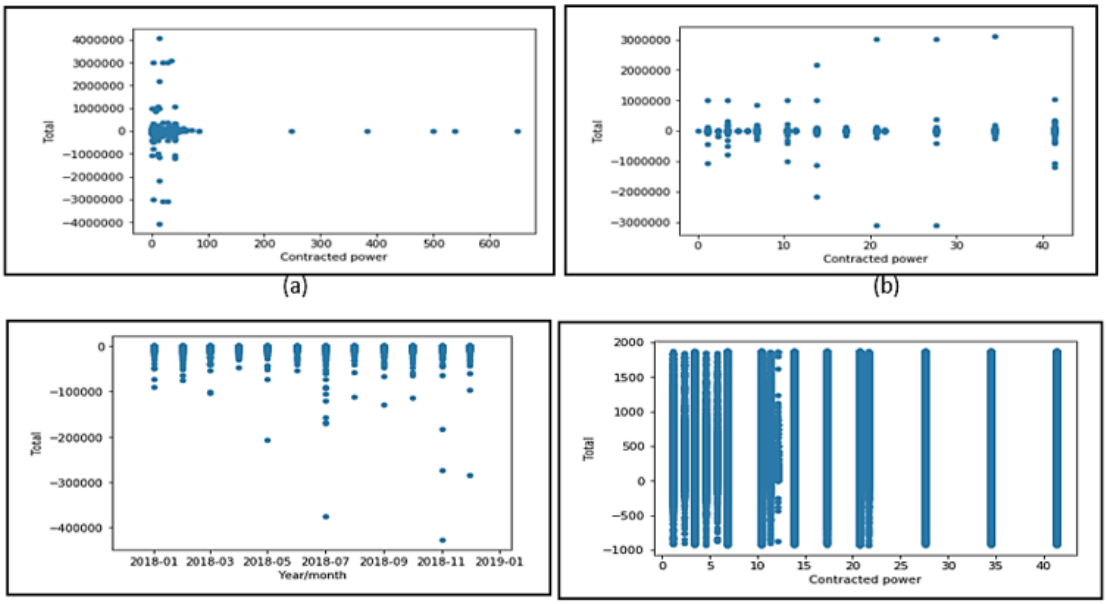


Figure 19. Data Preprocessing (Stage 1), (a) raw energy dataset (b) Public buildings have been removed that have several months less or more than 24 months, and public lighting buildings also have been removed because it is outside the scope of the study (c) public buildings that contain negative and zero values (d) outlier values have been removed using ISF, harmful and zero values have also been removed.

Figure 20 illustrates the process of stage 2 data pre-processing, wherein a subset of public buildings exhibiting negative or zero values is presented. In order to mitigate the occurrence of negative and zero readings, the technique of polynomial interpolation was employed to ascertain the compensation values for the data points (Xi, Yi), where Yi represents the energy consumption, the dependent variable, and Xi denotes the Year/month, the independent variable. Figure 20 depicts a representative example of a Portuguese public edifice, showcasing both zero and negative values, and elucidates the approach of employing polynomial interpolation to address this scenario. The proposed building exhibits a temporal representation wherein each point corresponds to a certain month spanning from the commencement of 2018 to the conclusion of 2019. Consequently, Figure 20 encompasses a total of 24 points. As an example, the data point corresponding to November 2018 is characterised by a value of zero, while the data point corresponding to September 2019 is characterised by a negative value. The interpolation method has been utilised to address the aforementioned issues in our public building dataset, including the resolution of zero and negative numbers.

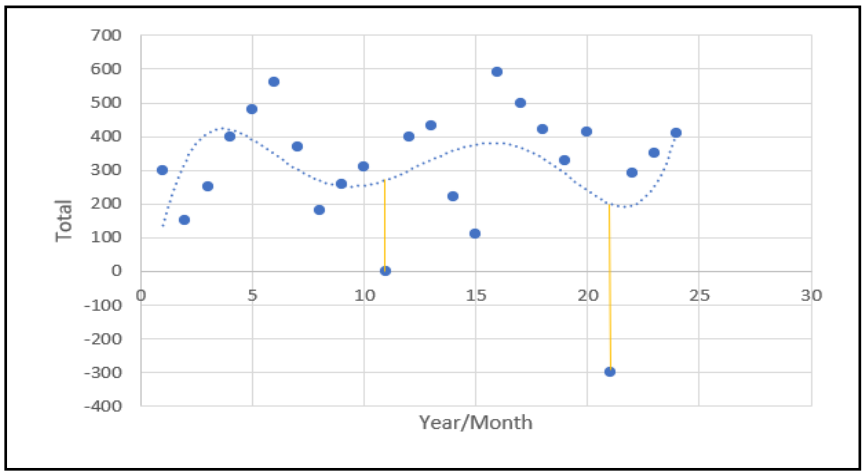


Figure 20. Data Preprocessing (Stage 2)

Each degree of the polynomial in the interpolation training was evaluated by its root-mean-squared error (RMSE), and the degree with the lowest RMSE was chosen for training. The degree-by-degree RMSE findings are displayed in Figure 21. Quintic polynomials have been used since they are considered the most reliable degree. There's no way to reduce overfitting by increasing the degree of polynomials.

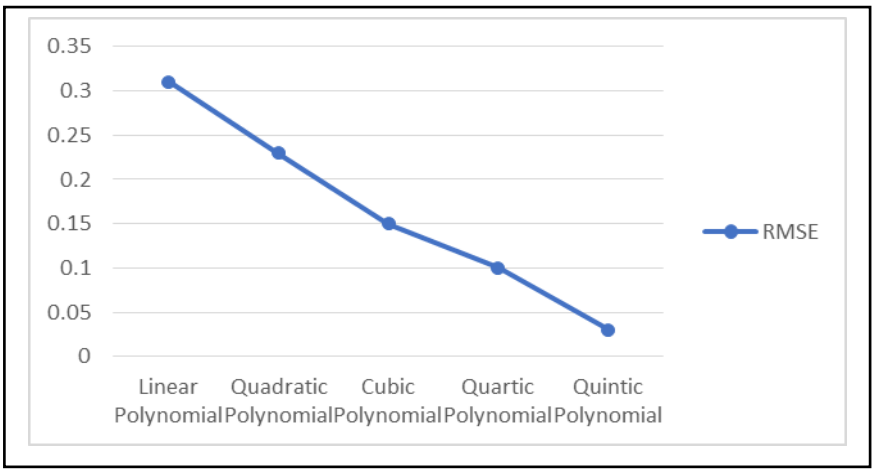


Figure 21. Degree of polynomials with RMSE

The procedure of data pre-processing resulted in the creation of the final dataset, which was subsequently utilised to discern discernible patterns in energy consumption across public buildings. Figure 22 illustrates a specific portion of the ultimate dataset, presenting the correlation between contracted power and overall energy use. It is evident that there is an absence of energy consumption numbers that are 0 or negative.

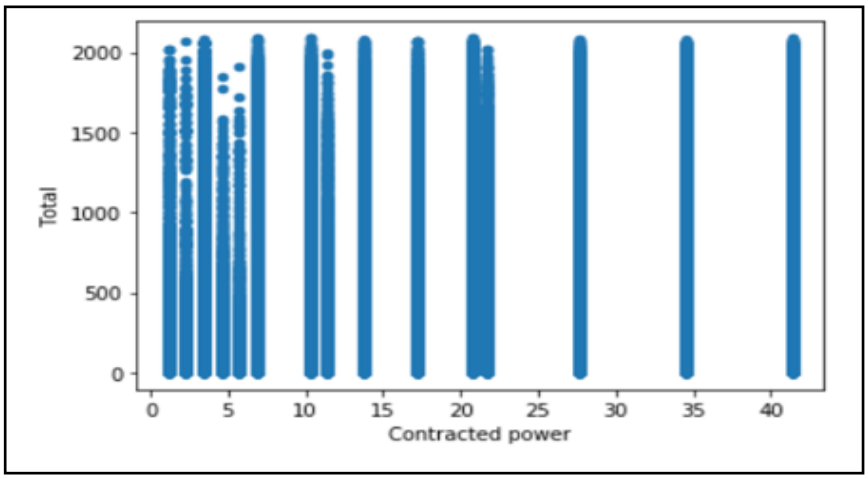


Figure 22. Total energy use versus contracted power

3.4.2 Results of Feature Selection

The aim of this section is to show the results of the T-test correlation coefficient and find the critical factors in the energy consumption dataset. Figure 23 shows the relationships between energy consumption factors. We observed a relationship between contracted power with Full, Peak, Empty, outside empty, and total consumption, and there is also a relationship between Full and Peak. Moreover, there is a relationship between Empty and Outside Empty. Moreover, we can avoid the Super Empty factor because it contains null values in all the columns, and there is no relationship between it and all the other factors. Finally, there is a negative relationship between the Simple factor with Full, Peak, and Empty consumption.

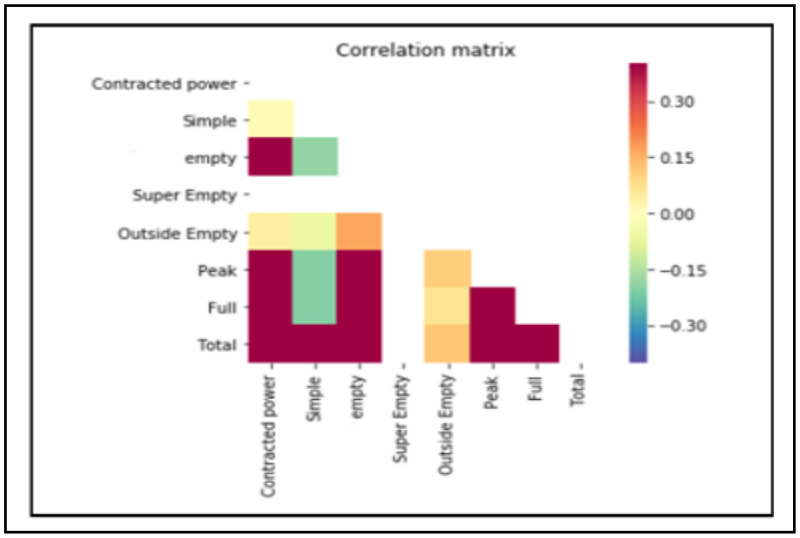


Figure 23. The applied correlation coefficient in the ECD.

3.4.3 Results of Finding Number of Clusters

This section presents the outcomes of three techniques employed to determine the most suitable number of clusters: the SOM, the Elbow approach, and the Bouldin and Davis method. A comparative analysis was conducted on the weights of the SOM network using two distinct approaches: one involving random weights and the other employing PCA weights. The number of iterations was set to 1000, while the values of sigma were set to 0.01, 0.1, 3, and 5. The comparison between random weights and PCA weights reveals that PCA weights exhibit superior performance in terms of q-error, particularly in the context of iteration = 1000 and sigma = 3. This finding is supported by the data presented in table 15, as well as Figures 24 and 25.

Table 15. A Comparison between Random Weights and PCA Weights

SOM random weights			SOM PCA weights		
Iteration	Sigma	q- error	Iteration	Sigma	q- error
1000	0.01	19.32	1000	0.01	348.24
	0.1	19.14		0.1	216.85
	3	18.02		**3**	**13.14**
	5	20.13		5	16.39

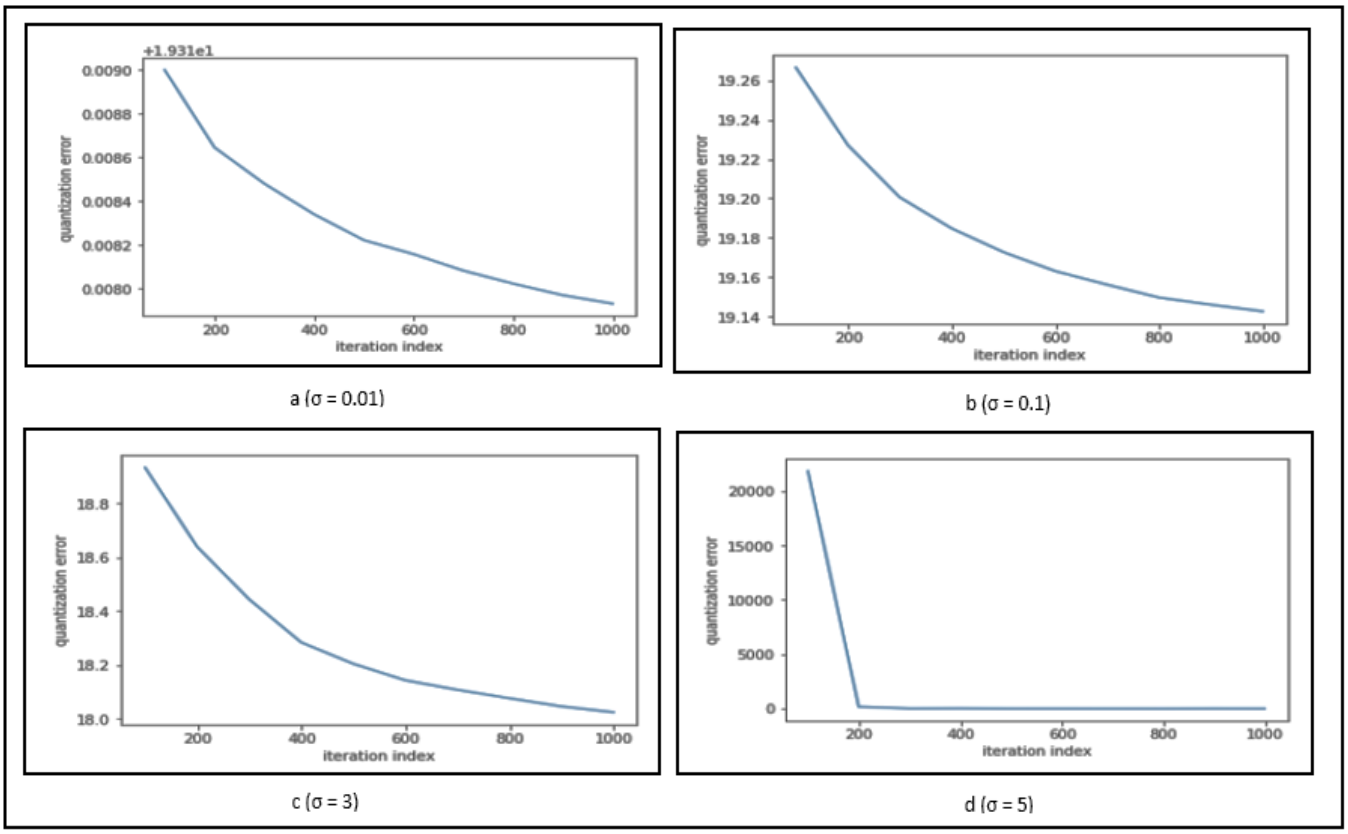


Figure 24. q-error in Random Weights. Set Iteration = 1000, (a) σ = 0.01, (b) σ = 0.1, (c) σ = 3 and (d) σ = 5

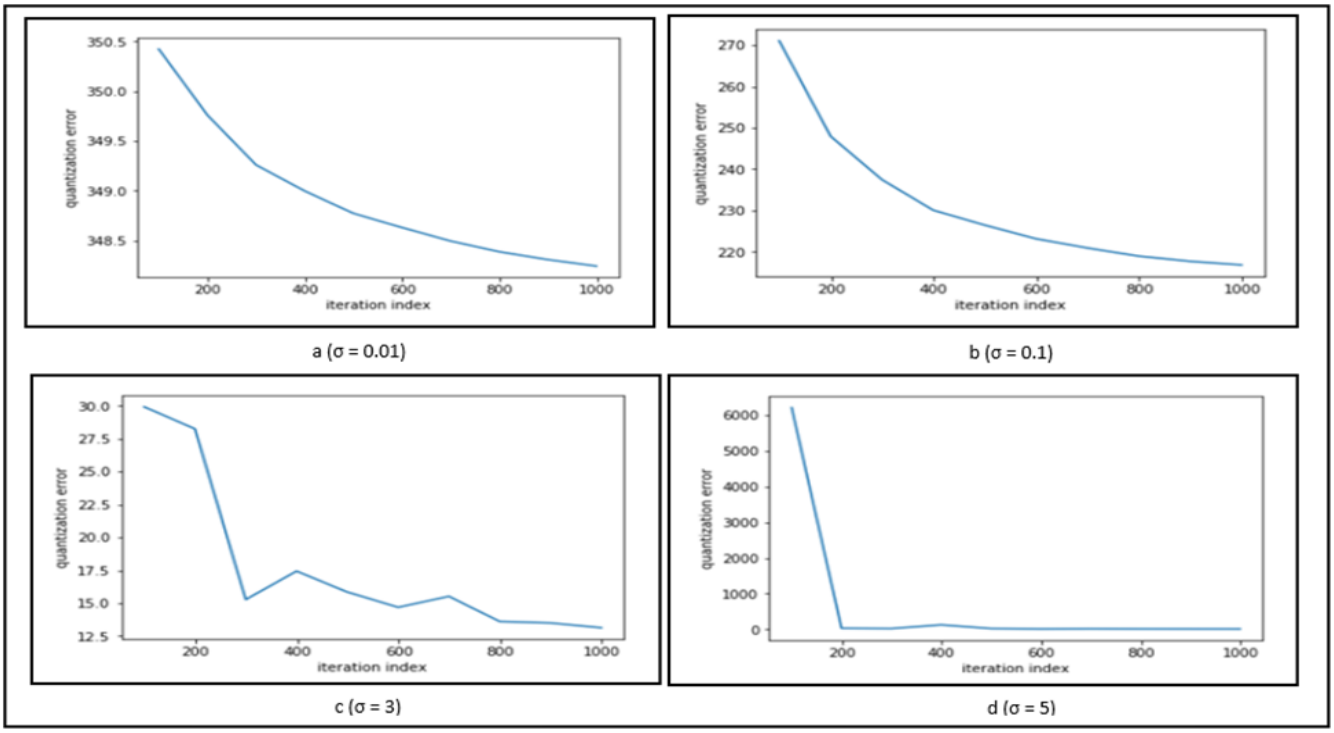


Figure 25. q-error in PCA Weights. Set Iteration = 1000, (a) σ = 0.01, (b) σ = 0.1, (c) σ = 3 and (d) σ = 5

The q- error expresses the squared distance (usually the average Euclidean distance) between input data x and their corresponding so-called BMU. Thus, the q-error reflects the average distance between each data vector (x) and its BMU, as shown in equation 17 (Müller, 2021):

$$\text{q} - \text{error} = 1/\text{N} \sum_{i=1}^{N} \left\| X_i - (BMU_{(i)}) \right\| \qquad (17)$$

The q- error appeared within table 15 and Figures 24 and 25 are midpoints for all data patterns. A comparative assessment of how this quantization is changed permits us to recognize distinctive clusters, which is one of the primary purposes of utilizing these techniques.

The SOM network was trained in two different ways based on PCA weights: random training SOM (RTSOM) and batch SOM (BSOM). The batch overhaul does not require a learning rate function. Typically, profitable since it reduces the number of required parameters. PCA weights with RTSOM (PCAW-RTSOM) are better than PCA weights with BSOM (PCAW-BSOM) in terms of q- error. q- error in PCAW-RTSOM and PCAW-BSOM is 8.97 and 9.24, respectively, as shown in table 16 and Figure 26.

Table 16. A Comparison Between PCAW-RTSOM and PCAW-BSOM

Iteration	PCAW-RTSOM	PCAW-BSOM
	q- error	
100	15.26	13.86
200	14.65	13.92
300	10.69	14.14
400	10.00	12.72
500	9.50	14.91
600	10.22	10.05

700	9.77	11.81
800	9.47	11.22
900	9.32	12.31
1000	**8.97**	**9.24**

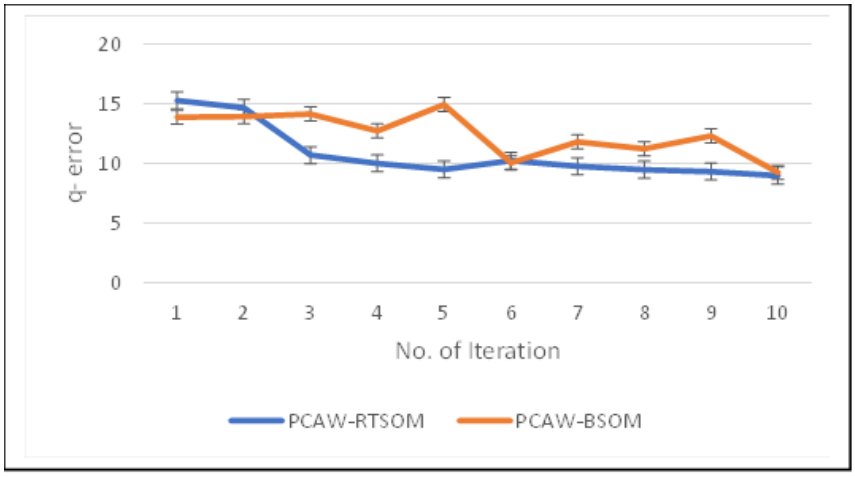
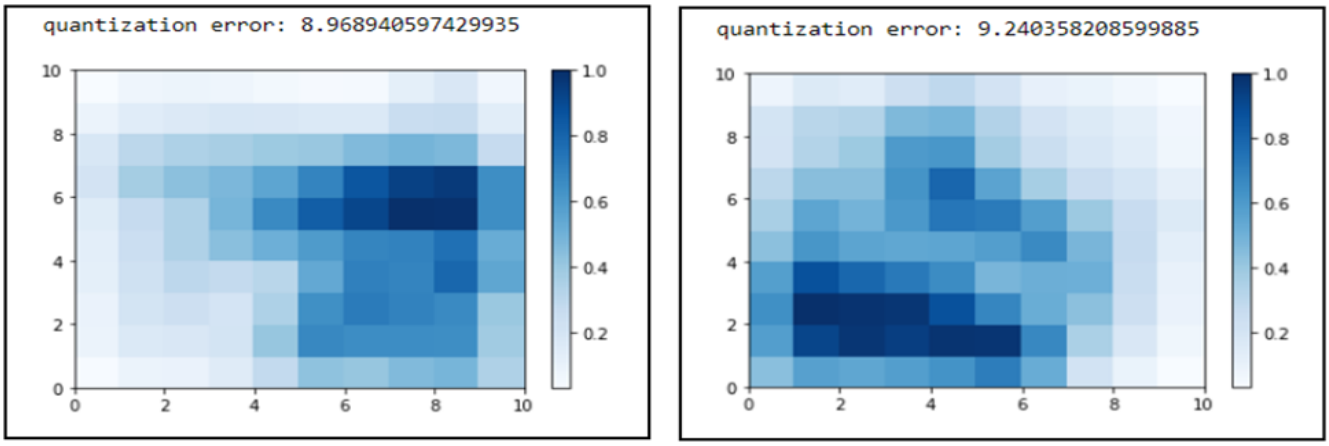


Figure 26. A Comparison Between PCAW-RTSOM and PCAW-BSOM

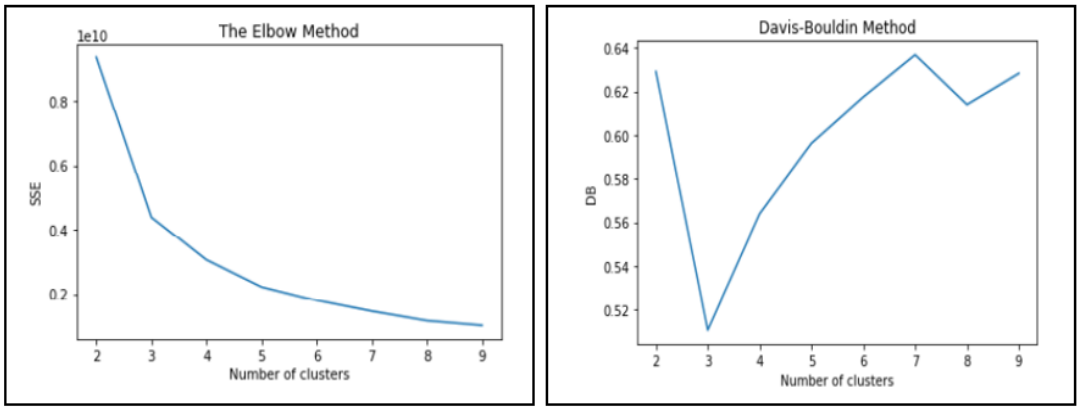


(a) (b)

Figure 27. U-matrix Comparison Between PCAW-RTSOM and PCAW-BSOM

Figure 27 displays the visual representation of the U-matrix in PCAW-RTSOM and PCAW-BSOM. Within the U-matrix, it is possible to identify three distinct regions of light colour, specifically white, which correspond to the lowest values in the U-matrix. These regions serve as indicators for three distinct clusters within the dataset on energy consumption. The aforementioned regions are demarcated by a dark blue hue, which corresponds to the delineation between the distinct clusters.

Three clusters were formed by applying the Elbow and Bouldin & Davis techniques to our ECD, as depicted in Figure 28.



(a) (b)

Figure 28. Comparison of the Elbow and Davis-Bouldin procedures on a dataset of energy consumption

The U-matrix exhibits dark regions that correspond to locations on the map where the points are significantly distant from one another, indicating the presence of segregation between the clusters. Conversely, lighter regions indicate fewer distances between points, suggesting a higher number of clusters. The Elbow technique is calculated by determining the sum of squared distances between the cluster centres and the data points within each cluster. The Euclidean formula is commonly employed in this context. The Bouldin & Davis approach involves calculating the intra-cluster scuttle by determining the average distance between each vector within a cluster and its centroid. This is achieved by employing the Euclidean method to measure the distance between the cluster's centroid and each vector. Ultimately, the ECD was subjected to analysis using the SOM network, Elbow technique, and Davis-Bouldin method, resulting in the identification of three distinct clusters categorised as low, medium, and high consumption.

3.4.4 KM with GA to Produce Energy Consumption Rules

In this section, the distance between each cluster was calculated using two methods: The initial method employed is KMC with KM++ initialization (KMCKI), while the subsequent method utilised is SPKG. The implementation of GA has been carried out using the primary parameters, as illustrated in Table 17. There exist three distinct methodologies for calculating distances between clusters, namely ED, MD, and CD. The performance of KMCKI and SPKG was compared in terms of SER, as shown by formula 18 (Ford & Siraj, 2013), and standard deviation. According to the findings presented in Table 18, the utilisation of CD with SPKG demonstrates superior performance compared to other approaches. Therefore, this study utilised a classification algorithm known as CD with SPKG to forecast cluster designations for each building within the ECD and identify latent patterns.

$$SE = \frac{\text{STDEV}(\Omega)}{\sqrt{\text{COUNT}(\Omega)}} \qquad (18)$$

Where:

- STDEV = Standard deviation
- Ω = Distances between each center of clusters

Visualising big data analytics is a crucial and significant undertaking. The clustering results have been demonstrated through several methodologies to assist decision-makers and stakeholders in the energy sector in Portugal in making informed choices regarding energy use in public buildings. Moreover, Energy Consumption Disaggregation (ECD) encompasses the noteworthy aspect of contracted power, which proves quite valuable in comprehending the varying energy consumption levels across different time periods within each public edifice. Figure 29 depicts a representative study of the diverse visual representations illustrating the dimensions employed in the ECD using CD with SPKG.

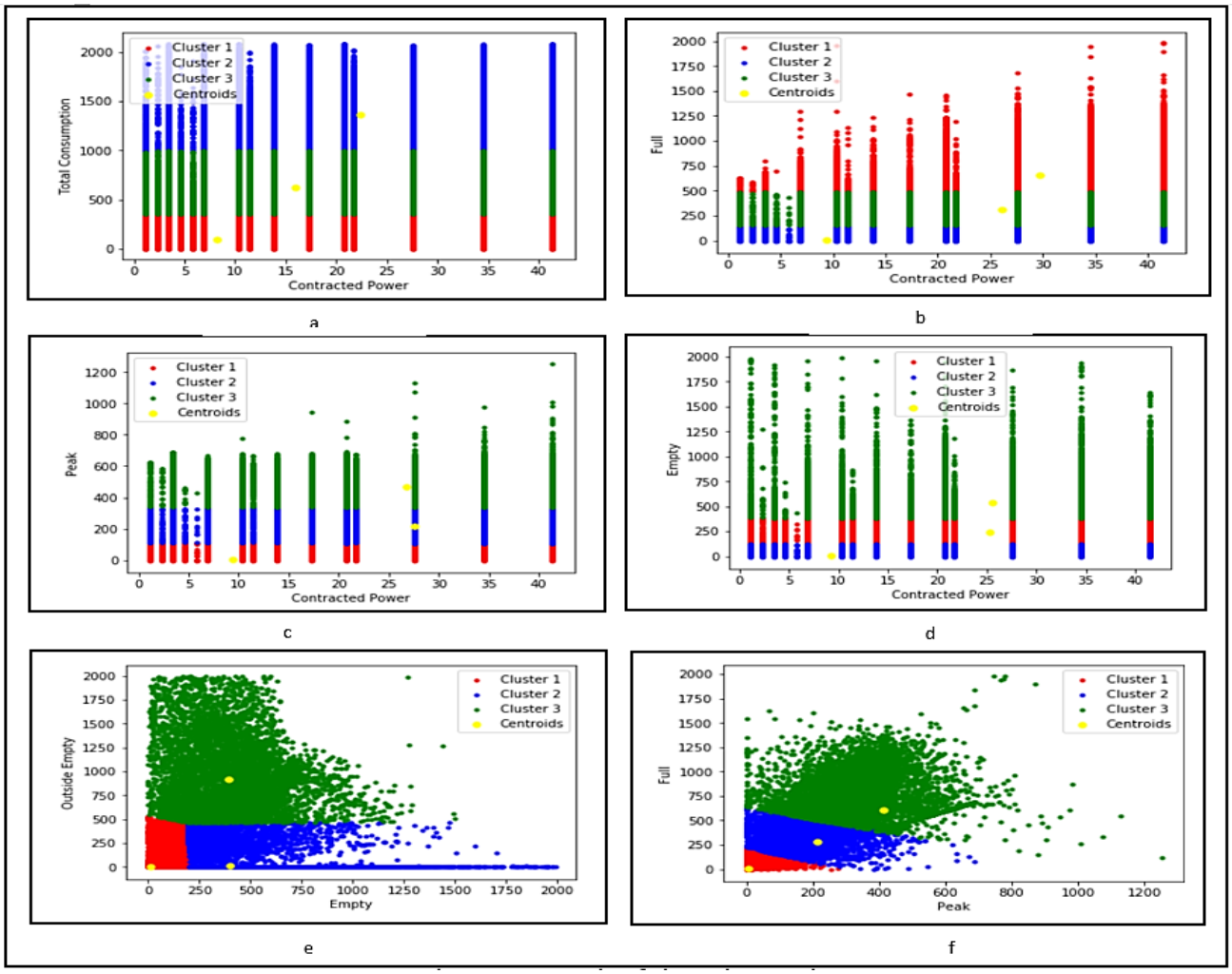


Figure 29. Sample of Clustering Results

Table 17. GA Parameters

No	Parameters	Value
1	Population Size	ECD
2	Crossover Probability	0.5
3	Crossover type	Two points
4	Mutation Probability	0.6
5	Mutation type	Bit flip
6	Number of Iterations	100

Table 18. A Comparison between KMCKI and SPKG in terms of SE and STDEV

No	Method	SE	STDEV
1	ED with Kmeans++ (EDK)	93.19	465.99
2	MD with Kmeans++ (MDK)	184.14	920.73
3	CD with Kmeans++ (CDK)	0.004	0.021
4	ED with SPKG	88.49	442.45
5	MD with SPKG	174.94	874.71
6	CD with SPKG	**0.002**	**0.012**

Through the examination of the clustering outcomes, a number of fundamental principles have been derived to aid stakeholders within the energy industry in Portugal in discerning the various typologies of public edifices, as depicted in Table 19. Energy consumption regulations assist decision-makers in identifying public buildings that require assistance for their occupants and facilitate the transition to alternative energy sources for such facilities. If-then rules have the potential to assist public buildings in managing their energy consumption through several mechanisms, such as:

- The if-then rules are characterised by their simplicity and accessibility, since they can be comprehended effortlessly even by individuals lacking expertise in the subject matter. As a consequence, both building managers and residents will experience enhanced comprehension of the impact of their decisions on energy use.
- The modifiability of if-then rules allows for their adaptation to suit the specific needs and demands of a given structure or organisation. This enables the implementation of energy management strategies that are more tailored and specific in nature.
- Real-time feedback can be provided to building managers and residents regarding their energy usage through the implementation of if-then rules. Individuals have the potential to modify their behaviour and make informed decisions on energy expenditure based on enhanced knowledge.
- The utilisation of if-then rules for energy management is considered cost-effective due to its ability to be applied using existing building management systems and sensors.
- One potential application of if-then rules is in the realm of energy conservation, wherein they can be employed to reduce energy consumption and thereby decrease associated expenses. The implementation of real-time feedback mechanisms and the promotion of energy-efficient behaviour can result in reduced energy consumption and financial savings for building managers and inhabitants.

Table 19. Sample of Energy Consumption Rules

No	Rules
1	Total<359 AND Full<157 Then cluster 1 (low energy consumption)
2	Total<359 AND Peak<111 Then cluster 1 (low energy consumption)
7	359<Total<992 AND 245<Outside empty<878 Then cluster 2 (medium energy consumption)
8	359<Total<992 AND 123<Empty<386 Then cluster 2 (medium energy consumption)
9	Total>=993 AND Full>=484 Then cluster 3 (high energy consumption)

10	Total>=993 AND Peak>=341 Then cluster 3 (high energy consumption)
14	157<Full<484 AND 111<Peak<341 Then cluster 2 (medium energy consumption)
15	Full>=484 AND Peak>=341 Then cluster 3 (high energy consumption)
16	Outside empty<245 AND Empty<123 Then cluster 1 (low energy consumption)
17	245< Outside empty <878 AND 123<Empty<386 Then cluster 2 (medium energy consumption)
18	Outside empty >=878 AND Empty>=386 Then cluster 3 (high energy consumption)
19	Total<359 AND Full<157 AND Peak<111 AND Outside empty<245 AND Empty<123 Then cluster 1 (low energy consumption)
21	Total>=993 AND Full>=484 AND Peak>=341 AND Outside empty>=878 AND Empty>=386 Then cluster 3 (high energy consumption)

Figure 29 exhibited superior performance in discerning energy consumption levels, although it was unable to ascertain the specific months characterised by heightened or diminished energy usage. Monthly consumption trends provide a comprehensive overview of energy use shown by occupants in public buildings. In the context of ECD, the energy consumption levels were established by utilising cluster label predictions, as depicted in Figure 30. The data presented in Figure 30 illustrates a discernible surge in electricity usage throughout the months of January, February, November, and December. Furthermore, there was a notable decline in energy consumption levels throughout the months of June and July. Additionally, this tool aids decision-makers in identifying the specific months during which energy consumption in public buildings has an increase. Consequently, the individuals residing in these structures are instantly directed.

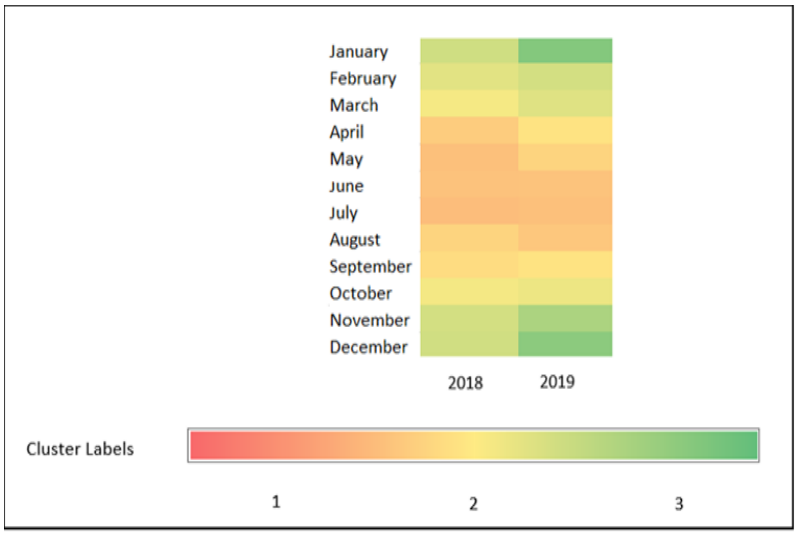


Figure 30. Monthly energy consumption patterns captured in different clusters for the ECD.

Figure 31 and Table 20 show municipalities and Portuguese public buildings activities that contain the number of buildings that consume low energy at different times. Three municipalities contain public buildings that consume little energy in figure 31, such as 'LOULE', 'SANTA MARIA DA FEIRA', and 'LISBON'. In addition, Table 20 shows Portuguese public buildings activities that consume little energy such as: 'INFRAESTRUTURAS PORTUGAL SA', 'GUARDA NACIONAL REPUBLICANA', and 'INSTITUTO SEGURANCA SOCIAL'.

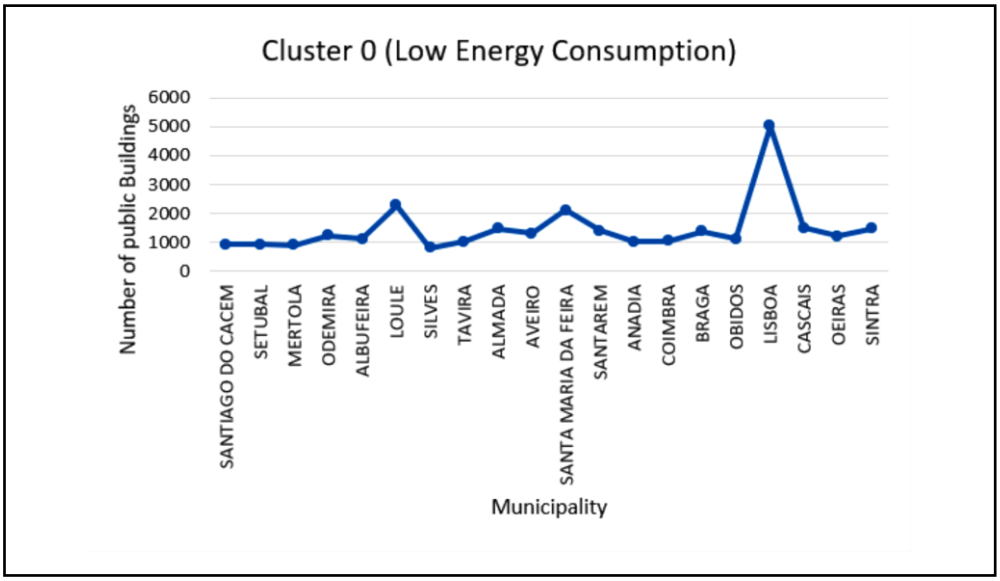


Figure 31. Sample of Municipalities that Consume Low Energy Consumption

Table 20. Sample of Public Buildings That Consume Low Energy in Each Municipality

Public buildings / Municipality	CACEM	SETUBAL	MERTOLA	ODEMIRA	ALBUFEIRA	LOULE	SILVES	TAVIRA	ALMADA	AVEIRO	MARIA FEIRA	SANTAREM	ANADIA	COIMBRA	BRAGA	OBIDOS	LISBOA	CASCAIS	OEIRAS	SINTRA
INFRAESTRUTURAS PORTUGAL SA	45	28	0	3	3	37	40	31	0	16	19	49	0	99	48	4	46	15	6	10
INSTITUTO SEGURANCA SOCIAL	3	0	21	7	6	12	12	0	13	0	23	0	0	0	0	0	0	0	62	5
ADMINISTRACAO REGIONAL SAUDE CENTRO IP	0	0	0	0	0	0	0	0	0	52	0	0	63	81	0	0	0	0	0	0
GUARDA NACIONAL REPUBLICANA	45	21	17	27	21	47	7	0	29	0	0	0	0	1	22	0	9	0	0	8

Table 21. Sample of Public Buildings That Consume Medium Energy in Each Municipality

Public building	Municipality	MONTIJO	CACEM	ALMODOV	MERTOLA	ODEMIRA	ALBUFEIRA	ALCOUTIM	CASTRO	LOULE	TAVIRA	MARIA FEIRA	MONTEMO	TORRES VEDRAS	BRAGANCA	VISEU	PORTO	VILA NOVA	LISBOA	CASCAIS	OEIRAS
IHRU INSTIT DA HABIT E REABILITACAO URBANA IP		0	105	0	0	0	0	0	0	23	0	0	0	0	0	0	2211	161	46	23	0
INFRAESTRUTURAS PORTUGAL SA		45	98	0	0	47	5	19	16	78	128	404	158	142	0	131	506	444	60	19	19
MUNICIPIO PORTO		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9797	0	0	0	0
MUNICIPIO OEIRAS		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8872

Table 22. Sample of Public Buildings That Consume High Energy in Each Municipality

Public building	Municipality	SINTRA	ODEMIRA	ALBUFEIRA	LOULE	SILVES	ALMADA	AVEIRO	SANTA MARIA DA FEIRA	AZEMEIS	COIMBRA	SOURE	GUIMARAES	BRAGA	BARCELOS	MIRANDELA	LEIRIA	VILA NOVA DE GAIA	CASCAIS	LISBOA	OEIRAS
GUARDA NACIONAL REPUBLICANA		1	31	22	39	20	31	0	0	0	0	0	0	20	1	18	0	29	0	24	0
ADMINISTRACAO REGIONAL SAUDE CENTRO IP		0	0	0	0	0	0	70	0	0	115	54	0	0	0	0	106	0	0	0	0
ADMINISTRACAO REGIONAL SAUDE NORTE		0	0	0	0	0	0	0	272	42	0	0	50	102	53	0	0	165	0	0	0
AUTORIDADE TRIBUTARIA E ADUANEIRA		11	7	0	19	0	14	0	28	0	0	0	3	0	2	18	0	0	13	23	0
INSTITUTO SEGURANCA SOCIAL		17	14	16	9	17	0	0	3	0	0	2	0	0	20	0	0	16	0	0	5

Figure 32 and Table 21 show the municipalities and Portuguese public buildings activities that contain the number of public buildings that consume energy on average between low and high consumption at different times. In figure 32, three municipalities contain public buildings that consume energy reasonably, such as 'PORTO', 'LISBOA', and 'OEIRAS'. In addition, Table 21 shows Portuguese public buildings activities that consume energy reasonably, such as: 'IHRU INSTIT DA HABIT E REABILITACAO URBANA IP', 'INFRAESTRUTURAS PORTUGAL SA', and 'MUNICIPIO PORTO'.

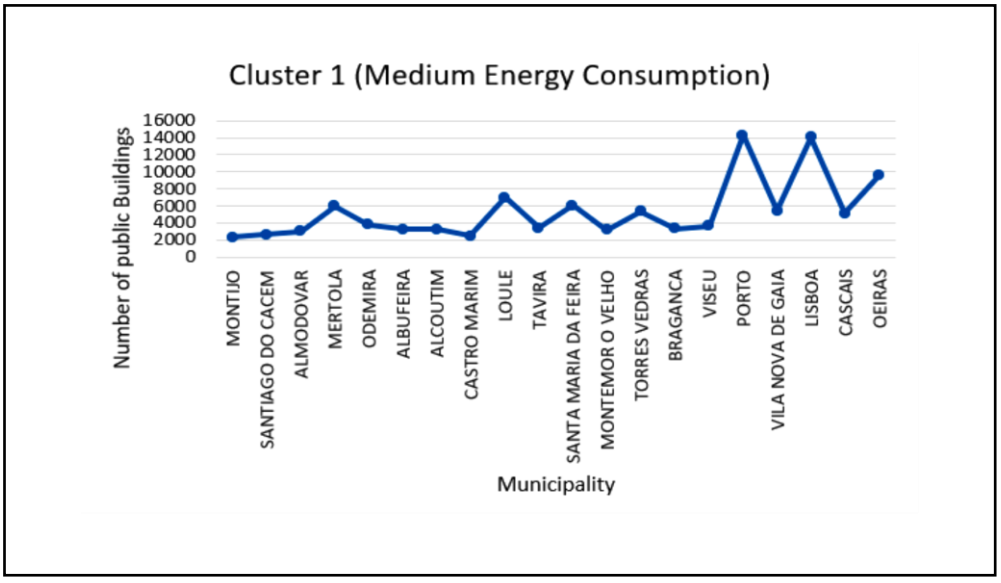


Figure 32. Sample of Municipalities that Consume Medium Energy Consumption

Figure 33 and Table 22 show the activities of municipalities and Portuguese public buildings containing the number of public buildings that consume high energy at different times. In Figure 33, four municipalities contain public buildings that consume high energy, such as: 'LOULE,' 'SANTA MARIA DA FEIRA,' 'VILA NOVA DE GAIA,' and 'LISBOA.' In addition, Table 22 shows Portuguese public buildings activities that consume high energy, such as: 'GUARDA NACIONAL REPUBLICANA,' 'ADMINISTRATION REGIONAL SAUDE CENTRO IP,' 'ADMINISTRATION REGIONAL SAUDE NORTE,' and 'AUTORIDADE TRIBUTARIA E ADUANEIRA.'

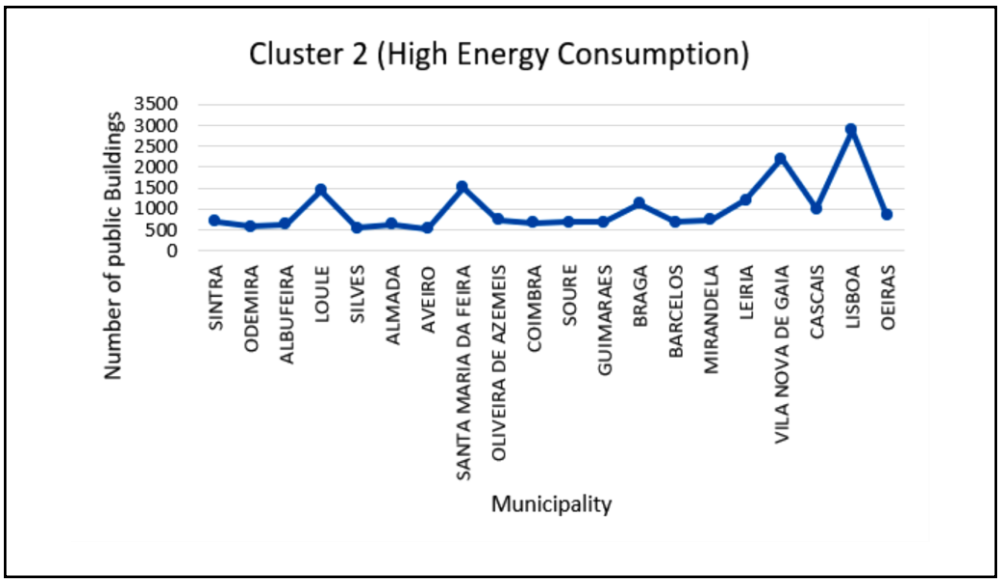


Figure 33. Sample of Municipalities that Consume High Energy Consumption

This precise research facilitates the identification of municipalities and Portuguese public buildings that require guidance for their users and a change in their energy providers.

By analyzing Figures 31 – 33 and Tables 20 - 22, municipalities such as 'LISBOA' and 'LOULE' contain public buildings with low, medium, and high energy consumption. In addition, there are Portuguese public buildings activities such as 'INFRAESTRUTURAS PORTUGAL SA' that consume low and medium energy. Therefore, we seek to find the distribution of the number of public buildings with different activities with low, medium, and high energy consumption over the different municipalities.

Tables 20, 21, and 22 show a sample of the public buildings located within each municipality. Knowing that each building has more than one location appears 24 times, distributed over 24 months over two years, 2018 and 2019.

By analyzing Figure 34, the number of public buildings in these Municipalities increased in certain months in 2018 and 2019 as follows:

- LOULE: Aug-18, Oct-18, Jan-19, Feb-19, Mar-19, and Oct-19.
- SANTA MARIA DA FEIRA: Feb-18, Mar-18, Apr-18, May-18, Jun-18, Nov-18, Jan-19, and Feb-19.
- BRAGA: May-18, Aug-18, Oct-18, Jan-19, Mar-19, Apr-19, and May-19.
- VILA NOVA DE GAIA: Aug-18, Sep-18, Oct-18, Nov-18, and Jan-19 to Oct-19.
- LISBOA: Feb-18 to Nov-18, and Jan-19 to Dec-19

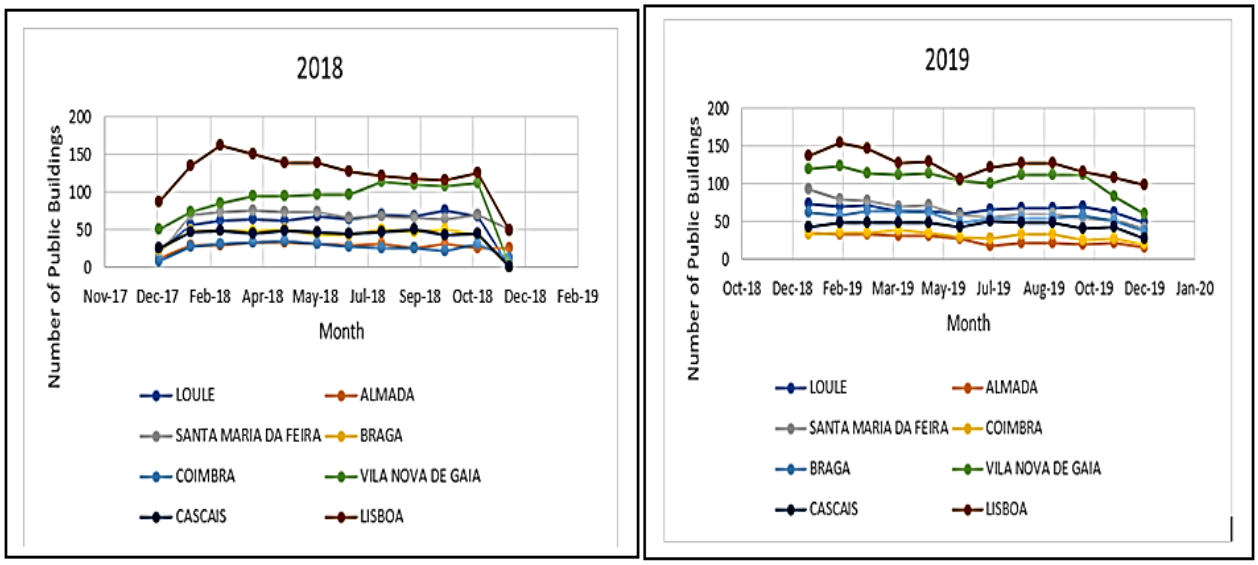


Figure 34. Sample of Public Buildings in Different Municipalities that Consume High Energy in 2018 and 2019

In relation to addressing our research inquiries, commencing with RQ1, which sought to gather data on energy consumption in public buildings in Portugal, and identify the key factors within this dataset that could assist in characterising said consumption, we successfully acquired aggregated monthly data for the years 2018 and 2019. This data pertains to a total of 77,996 buildings from diverse public sectors located in 238 cities across Portugal, resulting in a comprehensive dataset comprising 2,775,082 records. It has been determined that all components, or variables, within the acquired data are essential for the aforementioned profiling, with the exception of the variable that contains no data. The objective of our second research question (RQ2) was to identify the most suitable IC approaches for preparing the ECD in order to facilitate subsequent clustering analysis. In response to this inquiry, we employed many mathematical methodologies in order to achieve our objective. Specifically, we utilised outlier elimination through the implementation of the Isolation Forest algorithm, as well as polynomial interpolation approaches. In the context of a prepared dataset for

clustering analysis, our study aimed to address Research Question 3 (RQ3). The primary objective of RQ3 was to determine the optimal number of clusters within the ECD. To achieve this, we employed established techniques from the literature, including the SOM, the Elbow method, and the Davis-Bouldin method. Subsequently, our study aimed to propose a novel and optimised hybrid model for effectively classifying and labelling energy consumption patterns in buildings. The model utilised in this study, known as the SPKG model, incorporates a combination of various techniques, specifically SOM, PCA, KM, and GA. It was effectively employed to analyse our dataset and accurately predict the cluster label (low, medium, or high consumption) for each building. After obtaining a collection of buildings that have been appropriately identified, our focus shifted towards addressing Research Question 4 (RQ4). The objective of RQ4 was to identify significant patterns and overarching principles within this labelled dataset. These findings could then be utilised by decision-makers to optimise energy consumption in a logical and informed manner. Hence, an analysis was conducted on the clustering outcomes, leading to the formulation of a series of guidelines that facilitate the assessment of energy usage in a specific public edifice located in Portugal.

We have conducted a comparative analysis of our findings with the most advanced techniques documented in the relevant literature, specifically focusing on the utilisation of the KM algorithm. The study conducted by M. Azaza (Azaza & Wallin, 2017) and the study conducted by Al-Jarrah et al. (Al-Jarrah, Al-Hammadi, Yoo, & Muhaidat, 2017) are both relevant to the current discussion. The authors M. Azaza and Al-Jarrah reported SER of clustering as 28.3 and 22.5, respectively. The SER observed in our study is 0.002. Hence, our study demonstrates superior performance compared to earlier research by achieving higher ACC in the KM algorithm's.

3.5 Conclusion and Future Work

This study introduced an innovative hybrid intelligent model for the clustering of energy consumption levels (low, medium, high) in buildings. The model was evaluated using ECD in Portugal. In order to provide a conceptual framework for our study, we formulated four research inquiries that were adequately addressed. In order to get insight into the data, a correlation coefficient analysis was employed to identify the pivotal factors (variables) that impact the ECPB and ascertain the nature of the link between these factors. During the data preparation phase, ISF algorithm was employed to identify and eliminate outliers present in the dataset. Furthermore, an interpolation technique was employed to determine compensation values or approximate unknown values based on known values that are related. Regarding our modelling approach, our primary objective was to assign labels to the energy consumption level of each building. To achieve this, we initially determined the number of clusters of energy consumption present in the dataset. Through the utilisation of various techniques such as SOM, the Elbow method, and the Davis-Bouldin method, all three methods consistently identified 3 as the optimal number of clusters. These clusters corresponded to low, medium, and high energy consumption levels.

Subsequently, the KM was employed in conjunction with a GA to forecast the energy consumption cluster level for each individual building. This work makes contributions in four distinct areas. The initial study examines the various elements that impact the energy usage of buildings. The second approach presents an innovative framework for categorising the energy usage of public buildings into several tiers, such as low, medium, and high. The third study offers an analysis of extensive data regarding the ECPB in Portugal. Specifically, it focuses on data from the years 2018 and 2019, encompassing a total of 77,996 public buildings located in 238 cities around Portugal. For illustrative purposes, we successfully identified the municipalities characterised by high levels of energy use. Additionally, we have discovered the monthly energy consumption patterns of buildings for the years 2018 and 2019. The final component of the analysis involves deriving scientifically valid If-Then rules that aid decision-makers in rationalising energy consumption and identifying the public buildings with

the highest energy usage. This is accomplished by considering a set of three consumption levels (low, medium, or high).

Collectively, these findings can assist decision-makers in assessing the future energy needs of public buildings and promoting energy-efficient behaviours among residents.

As a suggestion for future research, alternative methodologies such as statistical approaches like multiple linear regression or logistic regression should be considered to identify key determinants that impact the energy usage of public buildings. One potential approach involves integrating the SOM algorithm with other optimisation approaches such as Grey Wolf Optimisation, Lion Optimisation, and Whale Optimisation. The objective of this combined approach is to determine the best quantity of clusters for analysing energy consumption data in buildings. Furthermore, the integration of clustering and optimisation methodologies, specifically grey wolf optimisation, lion optimisation, and whale optimisation, has the potential to enhance the ACC of cluster label prediction. In the context of forecasting the energy consumption of buildings, this research aligns with the prevailing literature by proposing the adoption of machine learning techniques, specifically deep learning methodologies such as LSTM, CNN, and deep forest.

Chapter 4 - Convolutional Neural Network with Genetic Algorithm for Predicting Energy Consumption in Public Buildings

A. Abdelaziz, V. Santos and M. S. Dias, "Convolutional Neural Network With Genetic Algorithm for Predicting Energy Consumption in Public Buildings," in IEEE Access, vol. 11, pp. 64049-64069, 2023, doi: 10.1109/ACCESS.2023.3284470.

Intelligent applications have become more important in the energy management of public buildings due to their ability to enhance energy consumption performance. Managing the energy consumption of these buildings poses a noteworthy challenge due to their erratic energy usage patterns and the absence of established design principles for enhancing energy efficiency and sustainability measures. Consequently, it is imperative to conduct an analysis of energy consumption patterns in public buildings and forecast forthcoming energy requirements. This evidence underscores the necessity of identifying and classifying energy use patterns in commercial and institutional buildings. The objective of this study is to determine the optimal intelligent approach for classifying and forecasting energy consumption patterns in buildings, focusing on public buildings as a case study. Additionally, the study aims to identify the scientific principles in the form of If-Then rules that can assist decision-makers in determining the appropriate energy consumption levels for each building. The objectives of this study were achieved through the utilisation of two IC models, namely the Elbow technique and the Davis and Boulden approach, for the purpose of quantifying the clusters of energy consumption patterns. The clustering problem was approached using both the KM algorithm and a GA. The evolutionary algorithm was employed to optimise the selection of centroid points for each cluster, hence enhancing the performance of the fitting model. The extraction of If-Then rules from cluster analysis has facilitated the process of identifying the buildings that have the highest energy consumption. CNNs augmented with GA were additionally utilised as intelligent models for the purpose of predicting energy usage. At this juncture, a GA was employed to optimise certain parameters of the CNN. The CNN model is surpassed in terms of ACC and SER measures by the CNN implemented with a GA. By employing a GA, the CNN attains a training dataset ACC of 99.01% and a validation dataset ACC of 97.74%. The corresponding SER for the training and validation datasets are 0.02 and 0.09, respectively. The CNN model demonstrates a high level of ACC, with 98.03% ACC with a SER of 0.05 on the training dataset. Similarly, on the validation dataset, the model obtains an ACC of 94.91% with a SER of 0.26. The research findings presented herein hold significant value for policymakers operating within the energy sector, as they furnish them with the necessary information to make well-informed decisions pertaining to the management of energy supply and demand specifically within public buildings.

4.1 Introduction

In the context of energy performance, buildings that exhibit inefficiency are identified as the principal contributors to global energy consumption and the subsequent release of greenhouse gas emissions (Pham, Ngo, Ha Truong, Huynh, & Truong, 2020). Consequently, it becomes imperative to prioritise the construction of buildings that exhibit reduced energy consumption and a more favourable environmental impact. The utilisation of energy in buildings has a significant contribution to the phenomenon of global warming, as well as the occurrence of air pollution and thermal pollution. These environmental issues have wide-ranging implications for human civilization (Cheng Li, Yang, Xiao, & Gao, 2023). The energy demand in public buildings has experienced a notable increase due to population growth and fast urbanisation in recent decades (Cai, Shen, Lin, Li, & Xiao, 2019).

The topic of energy efficiency in buildings has garnered significant attention from researchers, leading to the development of novel machine-learning applications (Fathi, Srinivasan, Fenner, & Fathi, 2020; Serale, Fiorentini, & Noussan, 2020; L. Li, Sun, Hu, & Sun, 2021; Nada & Hamed, 2019; Nordahl,

Boeva, Grahn, & Netz, 2019). Accurately estimating the energy requirements of a facility is of paramount importance in order to conserve resources and make well-informed decisions that ultimately lead to reduced energy consumption over time. Nevertheless, the task of forecasting energy consumption remains a formidable endeavour, mostly attributable to the multitude of elements that exert effect on this phenomenon. These factors encompass the physical characteristics of a structure and the energy consumption patterns exhibited by its occupants (Runge & Zmeureanu, 2019). According to Ding, Wang, Hu, and Wang (2022), ASHRAE, which stands for the American Society of Heating, Refrigerating, and Air-Conditioning Engineers, has categorised models for forecasting building energy use into two overarching categories: forward models and data-driven models.

Forward models, which are also referred to as physics-based modelling approaches, require multiple inputs related to the building and its environment. These inputs encompass the HVAC (Heating, Ventilation, and Air Conditioning) system, insulation thickness, thermal properties, internal occupancy loads, solar information, and other relevant factors (Z. Chen, Xiao, Guo, & Yan, 2023). DOE-2, Energy Plus, and TRNSYS are well recognised as prominent modelling tools that employ this particular methodology. These models require a multitude of parameters, many of which are frequently unattainable. Hence, the efficacy of these methods may be compromised due to their design complexity, computational demands, and limited data requirements (Pham et al., 2020).

In contrast, data-driven models exclusively depend on empirical data analysis. Several studies have put forth several models utilising machine learning techniques [(Pham et al., 2020), (Cai et al., 2019), (Z. Chen et al., 2023), (Helwig, Hong, & Hsiao-wecksler, 2020)] to estimate building energy consumption. These models have gained attention due to their ability to operate with little input requirements pertaining to the building's construction. The methodology is refined by utilising data extracted from BMSs (Building Management Systems) and smart metres in order to provide a substantial and all-encompassing dataset consisting of measurements recorded on an hourly or sub-hourly basis (Arjunan, Poolla, & Miller, 2022). According to Gouveia, Seixas, and Mestre (2017), the estimation of future building energy demand by machine learning models is influenced significantly by three key factors: the quantity of data available, the quality of the data, and the selection of an appropriate machine learning model.

According to several studies conducted by Li, Ding, Zhao, Yi, and Zhang (2017), Ouf, Gunay, and O'Brien (2019), Park and Son (2019), and Wen, Zhou, and Yang (2019), it has been suggested that the implementation of an energy modelling system with accurate predictions might potentially lead to a reduction in ECPB by approximately 10% to 30%. Consequently, it is imperative to sustain endeavours aimed at enhancing building energy prediction in order to foster the development of more energy-efficient structures. The advancement of data-driven models has resulted in precise energy estimates (Salam et al., 2020). According to McNeil, Karali, and Letschert (2019), the increase in greenhouse gas emissions, the proliferation of less energy-efficient buildings, the growing energy demand, and the need for energy savings will persist until the discovery of a dependable algorithm capable of accurately forecasting building energy use.

Several machine learning methods have been proposed in recent years for predicting future energy consumption in buildings [(Pham et al., 2020), (Cai et al., 2019), (Ding et al., 2022), (Z. Chen et al., 2023)]. Building energy use or consumption estimation has been facilitated by the utilisation of machine learning algorithms such as ANN, SVM, and Decision Trees. The majority of the data utilised for the training and evaluation of these algorithms is derived from datasets that encompass fewer than 1,000 buildings (Pham et al., 2020; Ruiz, Pegalajar, Arcucci, & Molina-Solana, 2020; Lee, Kim, & Ko, 2019; Nguyen & Aiello, 2013; Y. Chen, Chen, Yuan, Su, & Li, 2022). The existing literature consistently supports the notion that greater quantities of high-quality data contribute to enhanced precision in research outcomes. However, the restricted sample sizes of these datasets may

potentially yield model predictions that are less accurate (Chen, Guo, Chen, Chen, & Ji, 2022; Qavidel Fard, Zomorodian, & Korsavi, 2022; Chen, Dewi, Huang, & Caraka, 2020; Science, Bhuiyan, & Image, 2022).

The remarkable outcomes achieved by ANN have resulted in their growing prominence within the domain of energy forecasting. The literature extensively supports the notion that extensive and carefully maintained datasets furnish NNs with sufficient information for model training, hence conferring a notable advantage (Bourhnane et al., 2020). In their recent study, Fei, Chen, Liu, and Fang (2022) conducted an evaluation of the use of ANN in the domain of hourly building energy forecasting. The authors observed that the utilisation of the ANN algorithm demonstrated favourable outcomes in both single- and multi-step forward predictions. The performance of ANN in predicting energy consumption is evaluated using a dataset comprising two building blocks, each consisting of six stories. The energy modelling and simulation programme Energy Plus is employed for this analysis. According to Dong, Liu, Liu, Li, and Li (2021), the results of the study suggest that data-driven techniques, specifically ANN, exhibit superiority in the domain of building energy consumption prediction.

In order to predict the electricity consumption of a single hospital, the researchers examined the feasibility of utilising an ANN in conjunction with weather data and temporal fluctuations. According to the study conducted by Y. Chen, Chen, et al. (2022), the use of ANN forecasting showed superior performance during the colder months. Fu, Li, Zhang, and Xu (2015) were the pioneering researchers that introduced the concept of utilising a SVM for the purpose of forecasting building energy use. In the context of forecasting monthly electricity usage for four buildings using meteorological data, the researchers observed that the SVM exhibited superior performance compared to previous studies utilising NNs. The SVM model achieved a coefficient of determination (R2) exceeding 0.99. Chaowen et al. and Huang et al. employed a Bayesian NN methodology to forecast the hourly cooling requirement of a solitary office building. The RMSE outcomes for hourly load prediction utilising SVM demonstrated favourable results, as reported by Chaowen and Dong (2015) and Huang, Zuo, and Sohn (2016). Additionally, Dong et al. conducted a study in which they evaluated the performance of ANN and SVM in predicting the hourly energy consumption of office buildings. The study utilised a dataset consisting of 507 structures. The input variables of the model encompassed several factors, including dew point, air pressure, outdoor temperature, wind speed, and building data such as floor size and building type. As per the findings of Dong et al. (2021), the RMSE for the ANN model was recorded as 5.71, whereas the SVM model yielded an RMSE of 7.35.

The present study aims to investigate the application of optimisation approaches, such as GA, in order to enhance the ACC of clustering models and CNN models. The usage of a metaheuristic method is especially relevant in the context of addressing search and optimisation problems. The aforementioned process is characterised by the utilisation of one or more heuristics, hence acquiring the distinctive attributes of each heuristic employed. Consequently, a metaheuristic approach typically has limited empirical support for achieving convergence to the optimal solution, (i) demonstrates computing efficiency compared to exhaustive search, (ii) and aims to locate a solution that is close to optimal rather than the exact ideal solution (iii). The approaches employed in this context are iterative in nature and include the modification of one or more initial candidate solutions through the use of stochastic procedures, often created by random sampling of the search space.

The significance of accurate predictions in understanding the efficiency of building energy is widely acknowledged. However, existing literature research has not utilised datasets comprising more than 1000 buildings to enhance the predictive capabilities of models. Furthermore, none of the studies have achieved notable prediction performance according to the employed performance criterion. The authors, Runge and Zmeureanu (2019), have noted that the ACC of a model is highly dependent on the selection of the modelling technique, the quality of the data, and the quantity of the data. The

comparability of ACC results for algorithms implemented on different datasets is hindered by the presence of diverse data and varied situations, as these factors contribute to distinct outcomes [(Arjunan et al., 2022), (Luo et al., 2020), (Luo & Oyedele, 2021), (Divina, Torres, García-Torres, Martínez-álvarez, & Troncoso, 2020)]. Hence, it is imperative to conduct a comprehensive analysis and comparison of different modelling methods on a consistent dataset before making any claims about the superiority of one method over another. Several studies have conducted comparisons across different algorithms, utilising identical datasets, in order to determine the method that yields the highest level of precision in predicting energy usage within buildings. In order to determine the most precise methodology for attaining this objective, our study employs identical dataset and conditions across multiple modelling techniques.

The majority of research endeavours focus on analysing the aggregate energy consumption patterns exhibited by diverse architectural structures. Nevertheless, it is important to acknowledge that there are additional elements that need to be taken into account. These considerations include the consumption patterns of the individuals occupying the buildings, particularly during peak periods and vacant hours, which are specifically defined as the time intervals between 00:00 and 02:00, 06:00 and 08:00, and 22:00 and 00:00. This study introduces a cognitive computing model that autonomously categorises energy consumption into distinct tiers. In order to enhance energy efficiency and provide valuable insights for decision-makers regarding occupant behaviour in public buildings, we propose the utilisation of a hybrid intelligent model. This model integrates a CNN with a GA, enabling accurate predictions of energy consumption. The influence of our paper is diverse, covering four distinct areas.

- The proposed approach involves the integration of KM and GA to develop a novel hybrid model for the classification of building energy consumption into several levels, such as low, medium, and high. Furthermore, the optimal initial centroids in the KM algorithm are determined using GA.
- The proposed approach involves the integration of a CNN with GA in order to create a novel hybrid model for the estimation of building energy consumption. Furthermore, the parameters of CNNs are optimised by the utilisation of GAs.
- In this study, we utilised a comprehensive dataset encompassing the ECPB in Portugal. The information was collected over a two-year period, specifically in 2018 and 2019. Our primary objective was to employ this dataset for the purpose of training and testing our proposed models. Additionally, we aimed to evaluate the performance and ACC of these models. The dataset consisted of a substantial number of public buildings, totaling 81,260, which were distributed among 238 locations in Portugal.
- This study aims to present a cutting-edge intelligent model that can effectively analyse the energy consumption of buildings at the individual level, with the ultimate goal of enhancing energy efficiency in public buildings.

4.2 Related Work

The objective of this section is to furnish contextual information for our proposal and to justify the selection of strategies to be compared in the experimental section through a comprehensive assessment of pertinent prior research in the subject.

In their study, M. M. Ouf et al. employed a fusion of a Bidirectional LSTM (Bi-LSTM) network with a CNN in order to make predictions on building energy consumption. The discriminative feature values of the dataset were retrieved through the utilisation of a CNN, followed by making predictions utilising a Bidirectional LSTM (Bi-LSTM) network. K. Park and S. Son proposed a unique ensemble-based deep

learning model for the anticipation of energy usage and needs. Prior to being fed into the ensemble model, the dataset underwent typical pre-processing techniques, such as transformation, normalisation, and cleaning. Within the ensemble model, the CNN and Bi-LSTM network were employed to extract discriminative feature values. In order to optimise and ensure the ACC of the suggested model's predictions, an active learning strategy was devised in this research, employing the utilisation of a moving window. After this stage, the model was assessed on a dataset of Korean commercial buildings, utilising MAPE, RMSE, MAE, and MSE metrics to evaluate its effectiveness. In their study, Wen et al. (2019) employed a combination of the Extreme Learning Machine (ELM) and Variational Mode Decomposition (VMD) techniques to forecast forthcoming electrical loads. The VMD (Variational Mode Decomposition) technique was employed to divide the collected electric load time series into components characterised by different frequencies. This approach aimed to mitigate the influence of intrinsic fluctuations and enhance the overall predictability of the system. In conclusion, the utilisation of a differential evolution technique in conjunction with the Extreme Learning Machine (ELM) was employed for the purpose of forecasting.

In their study, Salam et al. (year) employed the integration of a Deep Belief Network (DBN) with linear regression techniques in order to provide predictions for time series data. The linear regression method is employed in this study to capture both the linear and nonlinear behaviours of the time series data. The disparity between the observed data and the expected data was initially assessed by linear regression, after which the DBN was provided with this value to inform its predictions. The Dynamic Bayesian Network (DBN) effectively distinguishes the distinct attributes pertaining to self-organization qualities and layers, rendering it a valuable tool for the purpose of time series forecasting. In their study on electric load forecasting, M.A. McNeil et al. employed SVM as a method for analysing time series data. The SVM method successfully captured the non-linear association between the target variables and exogenous factors. Various industries, including transportation, banking, aviation, and power/energy, have experienced significant advancements as a result of the utilisation of the innovative multivariate temporal convolutional network for time series prediction developed by Müller (2021). The convolutional network that was presented shown a significant improvement in the results of forecasting time series data. Furthermore, this study examines the trade-off between the ACC of forecasts and the level of complexity in the system. In their study, J. Lee et al. (2019) employed a GA in conjunction with a PSO technique to select optimal hyperparameters in the LSTM model, aiming to get the most precise outcomes in predicting building energy consumption.

In their study, Daim, Oliver, and Kim (2013) introduced a new oblique random forest classifier that may be utilised for time series forecasting. The proposed classification approach involves replacing each node of the decision tree with an orthogonal classifier that is deemed to be the most optimal given the supplied features. Furthermore, the process of feature partitioning was achieved by the utilisation of the least square classification approach. The efficacy of the oblique random forest classifier was examined by utilising a set of five electrical load time series datasets and eight general time series datasets. Furthermore, Chen et al. (2023) introduced a novel deep-learning network designed to predict forthcoming energy loads. The results obtained illustrate the robustness and strong ability to generalise of the deep energy model in the context of forecasting data series. Chen et al. (year) employed a Deep Belief Network (DBN) in conjunction with empirical mode decomposition (EMD) to forecast forthcoming electricity consumption. Prior to any additional analysis, the data series that were gathered were broken into many Intrinsic Mode Functions (IMFs). Subsequently, the DBN was employed to model each of the obtained IMFs in order to achieve accurate predictions. In their study, Wu, Huang, and Sutherland (2022) employed a random forest classifier to make predictions on energy consumption based on short-term energy consumption data. The efficacy of the random forest classifier was assessed on five datasets that covered a period of one year. The assessment results indicate that the random forest classifier discussed in this study had superior MAE in relation to its predictive ACC.

Qavidel. Fard et al. (2022) proposed the development of an ensemble classifier that utilises a combination of random forests, gradient-boosted trees, and decision trees to predict time series in extensive datasets. Upon undergoing rigorous testing, the constructed ensemble classifier demonstrated strong performance in the prediction of time series data. In their recent publication, Chen et al. (2020) proposed an innovative ensemble model that utilises stacking and multi-learning techniques for the prediction of time series data. The model described in this study integrates three main methodologies, namely SVM, linear regression, and a backpropagation NN. On the other hand, the ensemble model proposed in this research contains four fundamental processes, including integration, pruning, generation, and ensemble prediction tasks. In their study published in Science et al. (2022), researchers proposed a hybrid model that integrates a firefly algorithm with an Adaptive Neuro-Fuzzy Inference System (ANFIS) classifier for the purpose of predicting energy consumption. The incorporation of the firefly algorithm in this model leads to an increased diversity in the search space, hence enhancing the ACC of the predictive outcomes. S. Bourhnane et al. introduced a novel LSTM Multi-Seasonal Net (LSTM-MSNet) for time series forecasting that incorporates multiple seasonal patterns. The evaluation findings indicate that the LSTMMSNet model provided superior performance compared to existing approaches in terms of both computational efficiency and predictive ACC. In their recent study, Fei et al. (2022) proposed a methodology that combines multi-head attention with LSTM networks to enhance the ACC of time series data predictions. Z. Dong et al. (2021) initially eliminated outlier, redundant, and null values from the datasets through the utilisation of min-max and traditional transformation methods. Subsequently, an implementation of a Gated Recurrent Units (GRUs) model within a CNN architecture was employed for the purpose of energy consumption prediction. The experimental evaluation, using MAE, RMSE, and MSE, showcased the considerable performance of the offered model.

Efforts have been undertaken in the existing body of research to construct a computationally advanced framework for classifying the energy consumption of buildings based on diverse metrics that fluctuate with the temporal factors and the condition of the specific structure (Li et al., 2021). Stakeholders seeking to enhance the energy efficiency of buildings may find it advantageous to engage in the identification and categorization of energy load patterns exhibited by users in public buildings, utilising consumption profiles as a basis for this analysis. KMC is a technique that has been utilised in the research under evaluation. However, it does bring attention to certain issues. If the dataset being considered exhibits variations in volume and density, such as in the case of KMC, its effectiveness may be limited (Dong et al., 2021). Furthermore, it has been observed that the presence of outliers might lead to a displacement of centroids in a given dataset (Chen et al., 2023). In summary, the KM algorithm operates on the assumption that all variables possess equal variance, as stated by Chen et al. (2023), Arjunan et al. (2022), and Luo et al. (2020). Hence, the objective of our research is to enhance the efficacy of the KMC algorithm by identifying a more accurate method of clustering. Furthermore, previous research has demonstrated that the utilisation of big data in conjunction with a CNN has resulted in significant improvements in the prediction of building energy consumption. Hence, this study proposes a hybrid intelligent model, namely the CNN-GA, to predict future energy demands. In order to enhance precision and minimise the occurrence of MSE, the GA was employed to facilitate the training of the network, enabling it to ascertain the optimal weights during the training process. Hence, the proposed model's implementation by stakeholders in the energy sector enables them to make informed decisions regarding buildings with high energy consumption and to optimise energy supply for the occupants of such structures.

Through a comprehensive analysis of existing scholarly works, it has been ascertained that previous research endeavours have encountered difficulties in obtaining data that accurately represents the behavioural patterns of individuals within buildings over an extended period. Moreover, several studies utilise traditional clustering methodologies and statistical models, such as regression analysis,

without adequately assessing the effectiveness of these techniques in grouping energy consumption data or accurately predicting the levels of these groups. Inaccurate classification and prediction of energy consumption can lead to various forms of misguidance for decision-makers. These include the challenges of identifying buildings with high energy consumption, accurately anticipating the energy requirements of public buildings, and effectively identifying optimal energy providers. In order to bridge these deficiencies, a comprehensive dataset pertaining to energy use in public buildings in Portugal was created for the years 2018 and 2019. The information was utilised to train and evaluate classification and prediction models for public building energy usage. Our methodology has the potential to assist decision-makers in the energy industry in making informed decisions on the energy use of the public sector.

4.3 Research Questions and Methodology

The following questions were formulated to provide a framework for our investigation:

- Research Question 1: What are the potential data sources that can be utilised for the purpose of profiling a building's energy consumption?
- Research Question 2: Which IC methods may be utilised and adapted to ascertain the cluster size and cluster characterisation of the ECD?
- Research Question 3: What are the potential methods of IC that may be utilised and adapted for the purpose of clustering and predicting building energy consumption?
- Research Question 4: In the examination of energy consumption statistics, which trends have developed that are particularly noteworthy and distinctive?

The study proposes a hybrid methodology that integrates deep learning and optimisation techniques, specifically the combination of KM-GA and CNN-GA. This approach is suggested as a solution to the research questions posed in the study. The KM-GA and CNN-GA models are designed to cluster and predict energy consumption in buildings. The proposed methodology is demonstrated through a proof of concept conducted on public buildings in Portugal. Figure 35 provides a visual representation of the hybrid approach.

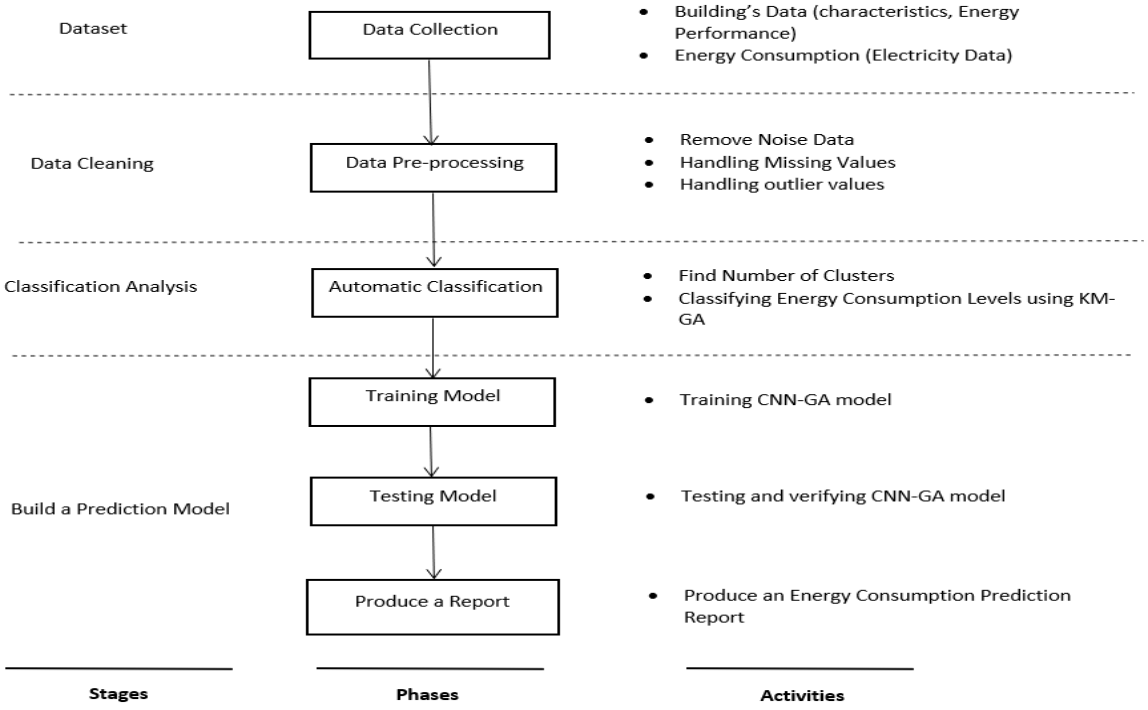


Figure 35. Our Suggested Model for Categorizing and Estimating Building Energy Use

The process of clustering and predicting building energy use is illustrated in Figure 35. The methodologically detailed approach will encompass the following stages:

1. We gathered data on variables like energy usage and building features, including but not limited to delivery point IDs, delivery addresses, contractual electrical power, electricity usage, and billing information broken down by month. The goal of this stage is to check that no significant changes were made to the underlying structure of the data, that all units of measurement are uniform, that sampling rates are sufficient, that the time series is stable over time, and that it is consistent with previous data. There are two parts, as follows:
   - The makeup, behavior, and energy efficiency of the building's data.
   - Use of resources (electricity data).

2. In this data preparation stage, we performed a thorough analysis of the data and, if necessary, changed it to disclose its information better. Outliers were removed using Isolation Forest (ISF), and missing values were filled in with polynomial interpolation. Three parts make up the whole process:
   - Clean up the data.
   - Treatment of missing data.
   - How to deal with extreme data (e.g., outliers).

3. Profiling energy consumption can be automatically sorted by categorization algorithms applied to the data. The Elbow approach and the Davis and Bouldin method were applied to our dataset to determine the number of clusters. After this process, all samples (rows) in the energy consumption dataset have had their consumption levels classified using our KM-GA technique.

4. Two deep learning models, CNNs and CNN-GAs, were used to train the data on energy use. These two models provided more precise and reliable energy usage predicting.

5. The two suggested models (CNN and CNN-GA) were tested for accuracy and error rate. We then picked the best model and proposed it to cluster energy consumption levels and assist stakeholders in making informed decisions about energy consumption.

6. Use the proposed methodology to build a report that predicts energy use, including which buildings will be more efficient and which will use more energy. As a result, this analysis aids decision-makers in the energy sector in three distinct ways:
   - Identify the highest energy-consuming buildings.
   - Estimating the energy consumption of future buildings
   - Assistance in switching energy providers, given the proper classification, and predicting the energy consumption of buildings.

4.3.1 Dataset and Clustering Preparation

Regarding data collection, data pre-processing, finding the number of clusters, and KM with GA sections, we have clarified them in Chapter 3, Section 3.3.1, Section 3.3.2, Section 3.3.4, and Section 3.3.5 respectively.

4.3.2 CNN with GA

CNN is classified under the realm of deep learning algorithms in the field of machine learning. The

format of CNN was designed with the human brain as a central consideration. The term "Convolutional" is derived from the utilisation of a linear mathematical operation called convolution, as opposed to solely performing matrix multiplications (Qavidel Fard et al., 2022). The grid-like architecture of the system has been well recognised for its effectiveness in managing data (Qavidel Fard et al., 2022). The dimensions of data processing vary from one dimension, which is applicable for processing signals and text, to three dimensions, which are suitable for processing images, audio, and video, and perhaps extend beyond.

In summary, CNN is composed of an input layer, an output layer, and a series of hidden layers which encompass several convolutional layers, normalisation operations, pooling operations, and fully connected layers. Typically, a convolutional layer is employed as the initial hidden layer, whereas a fully connected layer is utilised as the final layer. The convolutional layer is utilised to identify the associated qualities of the input data, while the pooling layer is responsible for combining the compared features. According to Li et al. (2021), when N represents the quantity of classes being categorised, the fully connected layer is responsible for converting the input into a vector comprising N dimensions. The basic architecture of a CNN is illustrated in Figure 36.

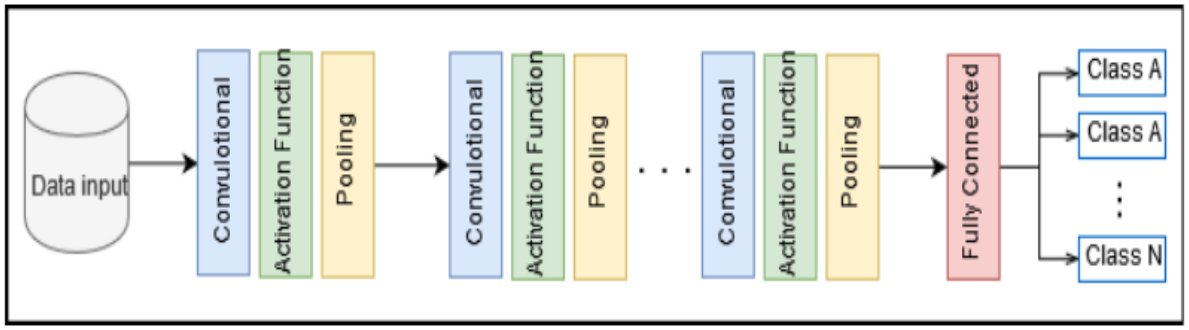


Figure 36. CNN Architecture

In the learning process, a loss function is employed to evaluate the classification's efficacy at each iteration. The network's prediction is evaluated by comparing it to the existing data, and a similarity metric is calculated. The concept of "Local Receptive Field" pertains to the observation that the neurons situated in the initial hidden layer of CNN are exclusively connected to a specific and limited segment of the input, known as the LRF. The weights and biases of any LRF model are initially created using a random process, as mentioned by Kim and Cho (2019).

The investigation focuses on the initialization of the weights. The output of each convolutional layer is subsequently passed through a nonlinear activation function in the subsequent layer. The purpose of this action is to enable CNN to ascertain the boundaries of its nonlinear decision-making processes (Kim & Cho, 2019). Various activation functions such as sigmoid, tanh, and ReLU are widely utilised and can be effectively employed in many scenarios. The outcomes can differ based on the specific activation function employed. The pooling layer is responsible for receiving the data after the activation function has been applied. The concept of pooling can be categorised into two types: maximal pooling and average pooling. Max pooling is a widely utilised technique (Kim & Cho, 2019) that involves selecting the maximum value within each patch of a feature map matrix. In contrast, the average pooling technique computes the average value over all the cells inside a given region. The outcomes of this calculation are presented in Figure 37. It is feasible to incorporate more convolutional layers; nevertheless, it is important to note that the ultimate layer consistently consists of a completely connected architecture, as depicted in Figure 36.

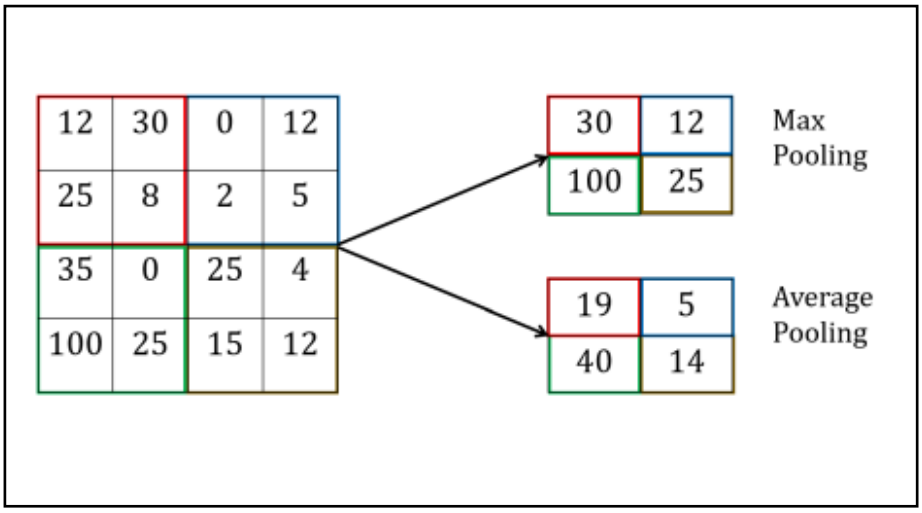


Figure 37. The Pooling Layer Types

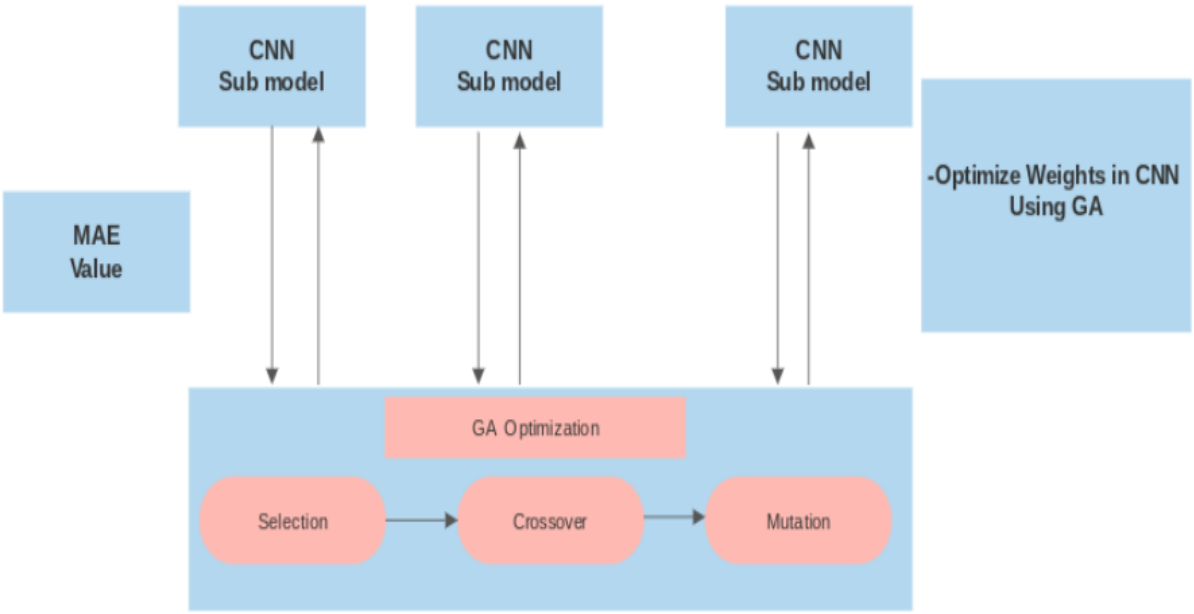


Figure 38. The Proposed Flowchart of the GA-enhanced CNN predictive model

The CNN classification model weights will be tuned with a GA optimization technique. While training the data, GA is used to determine the optimal model and weight values. To determine the classification accuracy, the training phase's best model is used to test data (see Figure 38).

The network weights are encoded in GA's chromosomes. The chromosomal count of the population is chosen at random. There are as many weight vectors as there are chromosomes. The training data's loss function (MAE) is the fitness function. Accordingly, when employing CNN, reducing the MAE of the training set becomes an optimization problem.

Python code that uses the fitness function to reduce loss has been developed. Eq. shows that the fitness value is the reciprocal of the loss value (19) (Kim & Cho, 2019).

$$\text{fitness value} = 1.0 / \text{loss} \qquad (19)$$

The following procedures determine the model's fitness:

- Get back the model's parameters from a one-dimensional vector.
- Indicate the values for the model's variables.
- In other words, guess what will happen.
- Determine the monetary worth of the damage.
- Find out how fit you are.
- Provide the fitness score.

With the release of PyGAD 2.8.0, a brand-new module known as Keras GA became available for use. Its full name, Keras Genetic Algorithm, is a mouthful, but the initials KGA suffice. Here are some of the features that may be accessed using the module:

Use the Keras GA class to construct a starting population of viable solutions. All Keras model parameters are available within each solution.

Utilize the model weights as vector () method to display the Keras model's settings as a 1-dimensional vector or chromosome.

The model weights as matrix () method in Keras may be used to get the model's parameters from the chromosome.

Keras GA class generates three instance characteristics in response to these two parameters:

- Reference to the Keras model.
- No solutions denote the total number of solutions in the population.

Population weights: A doubly linked list containing the model's parameters. When a new generation is created, this list is refreshed.

To get the best weight, use PyGAD and the code below to construct the fitness function. PyGAD's fitness function is a standard Python function with two parameters. The first stands in for the solution, while the second is the fitness value. Knowing where a solution ranks among the population might be helpful in some circumstances, which is why it appears as the second argument.

A 1-dimensional vector representing the solution is given to the fitness function. Step 1 demonstrates how to use the pygad.kerasga.model weights as a matrix() function to get the original Keras model parameters from the provided vector.

model = model, weights vector = solution, model weights matrix = pygad.kerasga.model weights as a matrix (Step 1)

In step 2, we see how the set weights () function updates the model to use the previously saved values for the parameters.

model.set weights (weights=model weights matrix) (Step 2)

Step 3 demonstrates how the model uses the predict () function to predict future results based on the current set of inputs.

model.predict(ECD) = predictions (Step 3)

The accuracy of the predicted outcomes determines the loss. As seen in Step 4, the MAE is employed as a loss function.

tens or flow. keras. losses. Mean Absolute Error () = mae (Step 4)

As indicated in Step 5, if the loss value is 0.0, it is best to add a small number, such as 0.00000001, to prevent a division by zero when determining the fitness value.

equation: solution fitness = 1.0 / (mae(data outputs, predictions)). NumPy() + 0.00000001) (Step 5)

After the GA has been run and the optimal weights have been obtained, the CNN is executed to predict the testing set of buildings' energy usage. CNN may be employed with only a single dimension. Nonlinearity is introduced via various layers, such as Convolutional layers, Pooling layers, Activation functions, and the Fully Connected layer. The Rectified Linear Unit (ReLU) activation function is used. Numerous studies [(Lee et al., 2019), (Kim & Cho, 2019)] have shown its efficacy and excellent accuracy in estimating energy use; hence it was selected. In addition, max pooling is used, and there is a 50% chance of drops. In addition, we use a maximum gradient of 5.0, a learning rate of 0.0005, and 100 epochs to train the model. Due to the one-dimensional nature of the ECD representation, a 1D Convolutional Layer is used.

This is the structure that is used:

- Embedding Layer
- 1D Convolutional Layer (Conv1D)
- Max Pooling Layer (MaxPooling1D)
- Relu Activation •
- 1D Convolutional Layer (Conv1D)
- Relu Activation
- Max Pooling Layer (MaxPooling1D)
- 1D Convolutional Layer (Conv1D)
- Relu Activation
- Max Pooling Layer (MaxPooling1D)
- Flatten Layer
- Dense Layer

Regarding testing and verifying the proposed model, 70% of the data is used for training, 15% for validation, and 15% for testing. The model is "trained" using the training data. Models are chosen based on the optimal solution (weight vector) that achieves the highest accuracy, as measured by the validation data. The suggested model is tested and evaluated with the help of the testing data. In addition, the accuracy and MAE are used to assess the proposed model, as indicated in Eq (20 and 21) (Kim & Cho, 2019).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (20)$$

$$MAE = \frac{1}{2}\sum_{i=0}^{n}|\theta - \Psi| \quad (21)$$

Where:

TP=how many predictions the classifier made where it correctly identified the positive class as positive.

TN = how many predictions the classifier made where it

correctly identified the negative class as unfavorable.

FP = the number of forecasts in which the classifier

erroneously forecasts a positive class for a negative class.

FN = how often the classifier misclassifies a positive class as a pessimistic prediction.

θ = the actual/true value

Ψ = the predicted/estimated value

4.4 Experimental Results and Discussion

This section is comprised of four sub-sections, namely: (1) data preparation, (2) cluster discovery, (3) categorization of ECD levels using KM with GA, and (4) prediction of ECD levels using CNN-GA.

4.4.1 Dataset Outputs and Clustering Results

Regarding data preparation, cluster discovery, and KM with GA classification of ECD levels sections, we have clarified them in Chapter 3, Section 3.4.1, Section 3.4.3, and Section 3.4.4, respectively.

4.4.2 CNN with GA to Predict Energy Consumption Levels

This section discusses two computational models that are utilised for the purpose of predicting energy usage. The initial weights of the first model (CNN) were not optimised during its development. In contrast, the second model, known as CNN-GA, was devised with the objective of optimising ACC and minimising the loss curve by the adjustment of the network's initial node weights. In order to determine the most efficacious model, we undertake a comparative analysis of these two models, focusing on their ACC and loss curve. This evaluation has the potential to assist stakeholders within the energy sector in making estimations regarding energy consumption levels.

The functioning of training and testing loss and ACC curves can be described as follows:

According to Figures 39 and 40, it can be observed that at epoch 100, CNN architecture attains the minimum training and validation ACC of 98.03% and 94.91% correspondingly. Additionally, the corresponding loss values are recorded as 0.05 and 0.26.

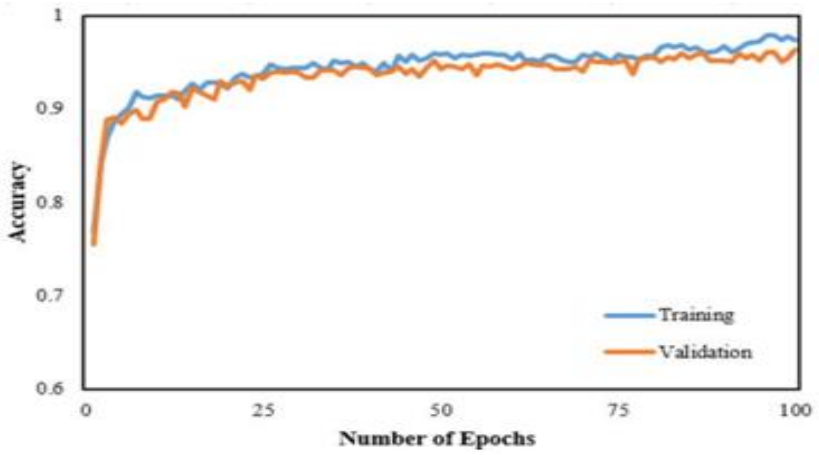


Figure 39. ACC of CNN Architecture

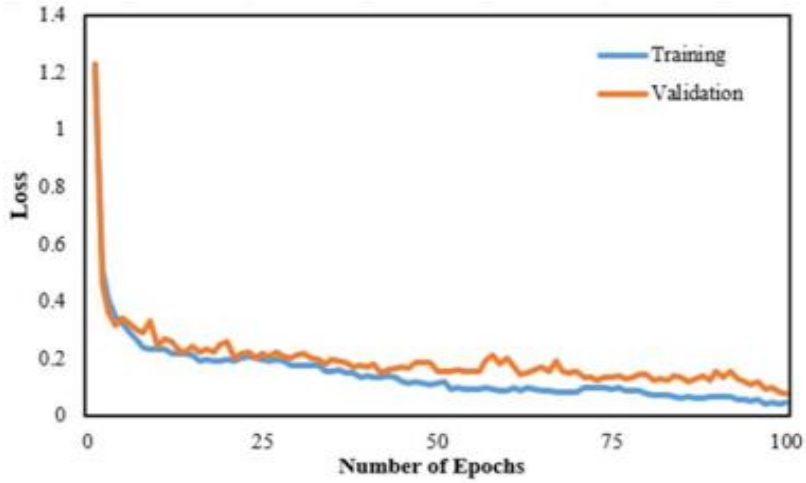


Figure 40. Loss of CNN Architecture

The results depicted in Figures 41 and 42 indicate that, around epoch 100, the CNN-GA architecture attains its highest levels of training and validation accuracies, reaching 99.01% and 97.74% respectively. Furthermore, the corresponding loss values are recorded as 0.02 and 0.09.

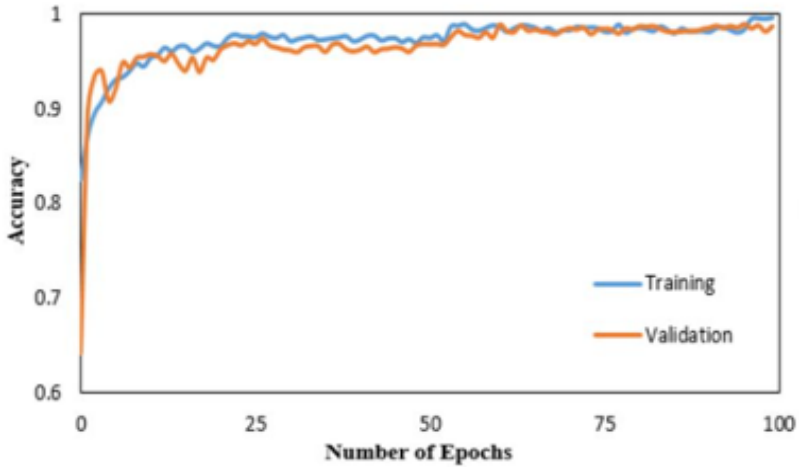


Figure 41. ACC of CNN-GA Architecture

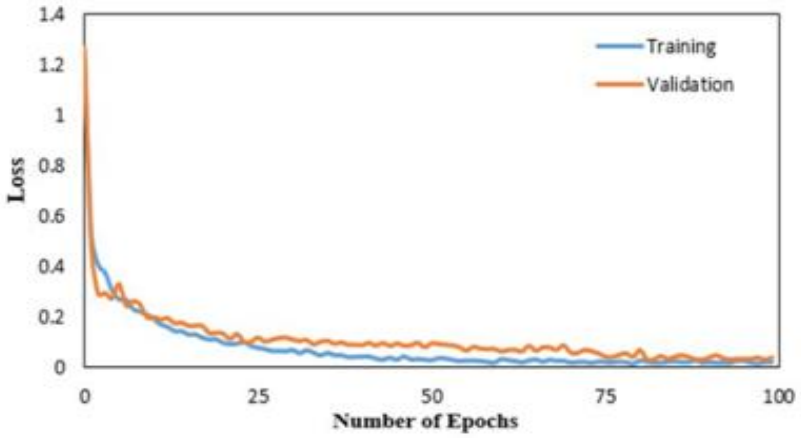


Figure 42. Loss of CNN-GA Architecture

Figures 43 and 44 clearly depict the predictive performance of CNN-based GA model for the dataset on energy usage. Upon conducting an analysis of these networks, it has been determined that the proposed CNN-based GA model demonstrates a high degree of ACC in predicting energy consumption.

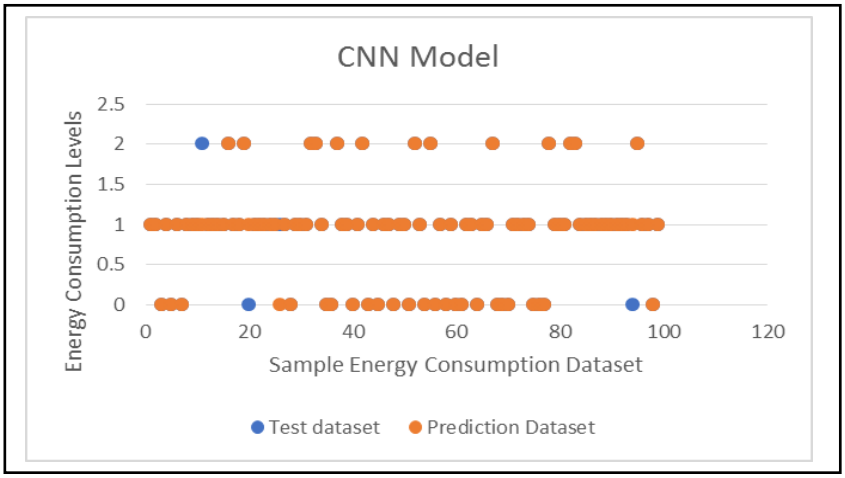


Figure 43. CNN Model Testing and Prediction in Energy Consumption Levels

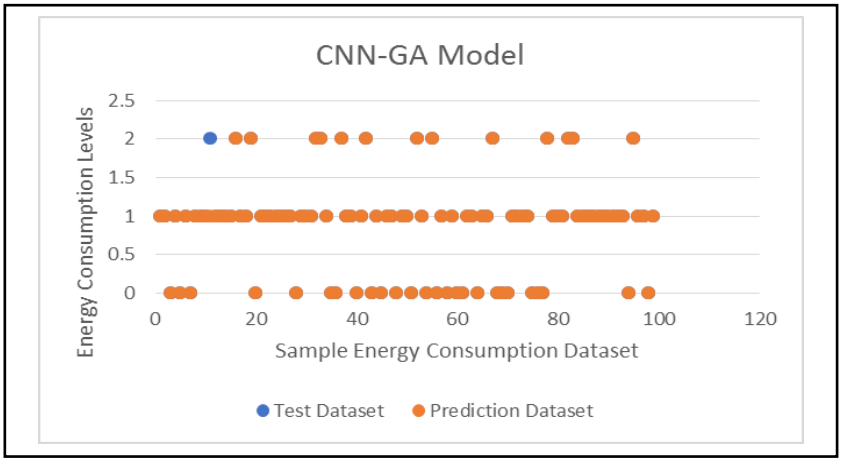


Figure 44. Evaluation of CNN-GA Models for Predicting Energy Use

4.4.3 Implications and practical applications in building energy consumption prediction

This article introduces a hybrid intelligence model that has been developed, trained, and validated using a dataset of energy consumption in Portuguese public buildings. The purpose of this model is to forecast future energy consumption in buildings. The derivation of essential rules was based on the analysis of energy consumption patterns from three distinct clusters, namely cluster 1, cluster 2, and cluster 3. In relation to energy policy, these rules provide assistance to decision-makers in the process of prioritising public buildings based on their energy use. This article additionally enabled the estimation of future energy usage in diverse public structures. Finally, the monthly trends in building energy use for the years 2018 and 2019 were computed. The findings presented in this study can be utilised by decision-makers to generate predictions regarding future energy requirements in various territorial dimensions. Additionally, these findings can aid in educating building occupants on the importance of efficient energy utilisation. Furthermore, decision-makers can make educated choices regarding energy providers with the assistance of decision support systems based on the aforementioned findings.

This research presents a novel approach to experimental design, proposing a building energy consumption prediction model that is both accurate and dependable. At the core of this model is an adaptive CNN. The energy-predicting approach being offered exhibits three notable advancements in comparison to prior deep learning models. These advances are as follows:

1) The energy consumption profiles of public buildings are categorised into many clusters using the KMC algorithm with GA. Subsequently, representative features are retrieved from each cluster. The optimal architectural design and weighting parameters for a certain CNN sub-model can be ascertained by leveraging the datasets inside each cluster.

2) Neurons within CNNs often establish connections with each other. Consequently, the comprehensive temporal correlation inside the dataset can be unveiled.

3) The utilisation of a GA has been chosen as a means to optimise the weights inside the hidden layers of a CNN. This selection aims to enhance the ACC of the suggested network and reduce the MAE measure when compared to the existing approaches employed in previous studies.

4) The dataset utilised for training and validating the proposed model for predicting energy consumption has a duration of two years, specifically 2018 and 2019. It comprises data obtained from a total of 26,624 public buildings located in Portugal. Precise energy consumption prediction is crucial for many processes, such as monthly building energy management, facility managers' decision-making, the creation of building information models, net-zero energy operation, and circular economy.

Accurate energy demand prediction plays a crucial role in the effective management of monthly building energy use. Accurate estimation of peak and monthly demand is crucial for the efficient scheduling and management of energy devices, hence enhancing the building's energy utilisation rate. The implementation of an accurate building energy consumption prediction can enable building managers to enhance their decision-making abilities in effectively managing various energy devices.

Also, accurate forecasts of energy usage could serve as the fundamental basis for effective smart energy management and the implementation of building energy efficiency retrofitting measures. Facility managers responsible for building management can get advantages from these predictions as they provide essential information for estimating forthcoming energy expenses and making informed decisions on the adoption of a more efficient pricing structure or the reduction of predicted energy use.

Moreover, performance-based building criteria are becoming more prevalent due to their ability to provide design flexibility while yet meeting or beyond the energy performance standards set by prescriptive-based requirements. However, it is imperative to utilise accurate building energy prediction models in order to assess the energy performance of a structure and determine whether the designs based on building information models can effectively accomplish the desired gains in energy efficiency.

Furthermore, the attainment of net-zero energy operation in buildings is contingent upon accurate energy forecasting. The global energy sector is undergoing a transformation as renewable energy sources steadily replace conventional fossil fuels. The determination of the output of active energy devices necessitates the precise estimation of energy demand and the incorporation of various renewable energy producing sources. Consequently, the coordination between the energy supply and demand of the structure can be optimised, facilitating its operation with minimal net energy expenditures.

Finally, an essential component in the establishment of a circular economy involves accurately forecasting the electricity demand placed upon the power system. The precise forecasting of power use has the potential to enhance societal and economic benefits through the reduction of energy usage and associated costs. In light of the persistent high energy demand exhibited by many structures, the reduction of energy consumption emerges as a crucial factor that could potentially influence economic growth.

The CNN-GA model, which has been previously evaluated in terms of the MAE metric, has been compared to existing state-of-the-art approaches in prior studies. The metric demonstrates that our proposed model exhibits superior performance compared to Adaptive LSTM NNs powered by GA (LSTMGA) (Luo & Oyedele, 2021) and GA-enhanced Adaptive Deep NN (GADNN) (Kim & Cho, 2019). In the suggested model, CNN-GA, the MAE achieves a value of 0.02 during the training phase and 0.09 during the testing phase. On the other hand, in the LSTMGA model, the corresponding MAE values are 0.51 during training and 1.15 during testing. Similarly, in the GADNN model, the MAE values are estimated as 0.63 during training and 1.71 during testing.

4.5 Conclusion and Future Work

This study introduces a hybrid intelligent model to anticipate future energy consumption levels in public buildings. The model is designed to address the four research objectives identified for investigation. Between the years 2018 and 2019, a comprehensive collection of raw data was conducted on a monthly basis. This data was obtained from a total of 77,996 buildings, encompassing various public sectors, and spanning across 238 cities within the country of Portugal. An isolation forest algorithm was utilised to identify and remove outlier values present in our energy dataset. Additionally, interpolation techniques were employed to determine compensation values or estimate unknown values by using the information provided by related known values. Following the completion of data preparation, the total count of records utilised in this study amounted to 1,222,695. These records corresponded to a total of 26,624 public buildings. It is important to note that records pertaining to public lighting were excluded from the analysis, as they fell outside the scope of our investigation. Additionally, buildings lacking consumption data for the entire 24-month observation period were also excluded from the dataset.

The application of GA was employed to implement the KM algorithm for the purpose of predicting cluster labels within each building. Simultaneously, the Elbow technique and the Davis and Boulden strategy were utilised in order to ascertain the optimal number of clusters. Moreover, the utilisation of a GA has facilitated the derivation of If-Then rules from KM data, hence assisting in the identification of buildings characterised by the highest energy consumption. Two IC systems, namely CNN and GA

based CNN-GA, have been created for the purpose of predicting future energy use. This discovery has implications in four distinct domains. The initial proposal introduces a novel methodology for categorising the projected energy usage of public structures into distinct tiers, such as low, medium, and high. The second dataset comprises a substantial volume of data pertaining to the energy use of government buildings in Portugal during a span of two years. The third component entails the extraction of scientifically solid If-Then principles that may be utilised by decision-makers to justify the ECPB and identify the most significant contributors to energy usage among them. In this study, two sophisticated models have been employed to create prognostications on forthcoming energy consumption, while also evaluating the precision and standard deviation of these models.

Suggestions for future research encompass the utilisation of statistical methodologies such as multiple linear regression or logistic regression to ascertain significant factors that impact public building energy consumption. Additionally, the amalgamation of clustering and optimisation techniques, specifically grey wolf, lion, and whale optimisation, could be employed to enhance the precision of predictions pertaining to cluster labels that describe building energy consumption levels, namely lo The present study also suggests the utilisation of machine learning algorithms belonging to the deep learning category, such as LSTM and deep forest, as a means to enhance the precision of forecasts pertaining to the energy consumption of a structure.

Chapter 5 – Discussion

This section will elucidate the primary discoveries and ramifications of our investigation on energy usage in public buildings through clustering and deep learning methodologies.

5.1 Key Findings

In the first section, we pinpoint the essential elements that affect how much energy is used by public buildings, pinpoint the most famous intelligence computing methods, cluster and predict the energy consumption in those buildings, and pinpoint the performance metrics that have been used in the previous work in these situations.

We may highlight the key findings from our thorough investigation. When examining the energy use of buildings, the phrases "electricity," "heating," and "climate" appear to be the most pertinent. Also, we may get the conclusion that "KMC" is significant when using in-depth analysis in the earlier study. In the context of predicting energy consumption in buildings, the phrases "backpropagation," "feed-forward NN," and regression models like "multiple linear regression" and "SVM" are also crucial. Hybrid techniques like KMC with backpropagation and feedforward NNs have been explored for clustering and predicting energy usage in buildings. Finally, the ACC and MSE were the most crucial metrics to assess the performance of the intelligent model, according to the previous work's analysis of a variety of performance evaluation metrics.

In section 2 and 3, we focused on the electricity factor that extracted from the previous works based on it is important factor that use for clustering and predicting energy consumption in buildings. Regarding other important factors: heating and weather, we will use them with electricity in the future to improve our intelligent model. In this study, the energy used in public buildings in Portugal that have the following characteristics makes up the data for this study: Data was gathered monthly between the years of 2018 and 2019 in 238 cities, 77 996 buildings, and other public sectors, totalling 2 775 082 records. The number of records used in this study totalled 1 222 695, or 26 624 public buildings, after removing the records pertaining to public lighting (since it is outside the purview of our study) and buildings that did not contain consumption data for the full observed period of 24 months.

The isolation forest and interpolation methods have been utilized in dataset preprocessing to clean our dataset so that our intelligent model may be built with high ACC. Outliers have been identified using the isolation forest and missing and negative data have been compensated using the interpolation method.

In section 2, we focused on to get the important results such as extract the critical features from our energy dataset by using coefficient analysis, also identify the number of clusters in our dataset by using many techniques for instance SOM, Elbow method, Davis& Bouldin method, and predict cluster labels by hybrid model which is GA with KMC. The cluster labels have been divided into three main clusters (Low, Medium, and High energy consumption). In such a tagged dataset, we find key trends and overarching principles that could aid the decision-maker in rationalizing energy consumption. As a result, we examined the clustering data and developed a set of guidelines that can aid in characterizing the energy usage of a specific public structure in Portugal. Finally, determine the municipalities and portguess buildings activities that high consume energy consumption for instance "GUARDA NACIONAL REPUBLICANA" includes 24 public buildings in "Lisboa" , "ADMINISTRACAO REGIONAL SAUDE NORTE" includes 272 public buildings in "SANTA MARIA DA FEIRA" and others.

In section 3, we focused on extracting three main points, as follows:

Creating a special hybrid model by combining KM with a GA to classify building energy consumption into levels (such as low, medium, and high). In addition, with the aid of GA, the optimal starting centroids in KM are found, as clarify this point in section two.

Creating a special hybrid model for predicting building energy level by fusing CNN with GA. Additionally, GA is used to optimize the CNN settings.

Recommending a cutting-edge intelligent methodology for assessing building-level energy use to enhance the energy effectiveness of public buildings.

Our study aimed to investigate the correlation between energy consumption in buildings and the application of clustering and deep learning methodologies. Several significant discoveries were discerned through the examination of a dataset encompassing energy consumption trends across several buildings.

5.1.1. Identifying Critical Factors in Energy Consumption

The identification of crucial common characteristics pertaining to energy consumption in buildings is conducted by a comprehensive literature evaluation in this work, utilizing text mining and bibliometric mapping techniques. The crucial shared element is electricity, as depicted in the figure 45.

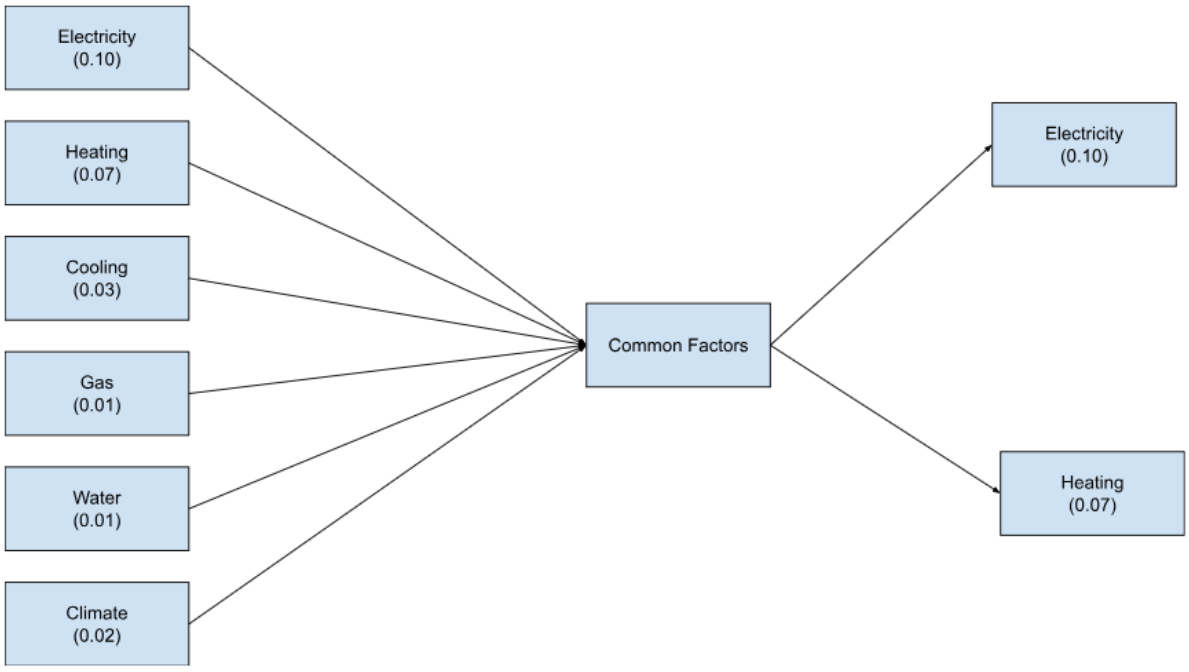


Figure 45. Determine the Common Factors in the Literature Review

Note: if (TITs) > 0.05, the term is relevant.

This study employs many features related to electrical factors for grouping and predicting energy consumption in public buildings. The electrician was relied upon at different times throughout the day. This is for two main reasons. First, most previous studies relied on this main factor because it is the most common among decision-makers in the energy field and the focus of attention of agencies and countries interested in reducing the energy consumed in buildings. Second, the available data obtained

from the Adene Energy Agency (This agency is in Portugal) were explicitly based on the electricity factor.

Correlation analysis has been employed to identify the crucial features pertaining to the electricity factors, as depicted in the accompanying Figure 23.

Furthermore, the existing literature has established that intelligent algorithms are often employed for clustering and predicting the energy consumption of buildings, including KM and deep neural networks such as CNN and LSTM.

5.1.2. Identifying Energy Consumption Patterns

We successfully categorized buildings into distinct clusters by employing clustering methodologies according to their energy use trends. This study provided valuable insights into the various aspects that contribute to energy efficiency or wastage in buildings.

A novel hybrid model, known as SPKG, was employed to cluster the energy consumption of public buildings into distinct levels. The model demonstrated high accuracy in terms of standard error and standard deviation, thus achieving the clustering task. In order to ensure the precision of clustering, two hybrid techniques were employed, namely K-means clustering with K-means++ initialization (KMCKI) and K-means with GA (KMGA), to ascertain the optimal centroid for the clustering process. The strategies were evaluated and the most optimal one was chosen based on the evaluation of standard error and standard deviation, as depicted in Figure 46.

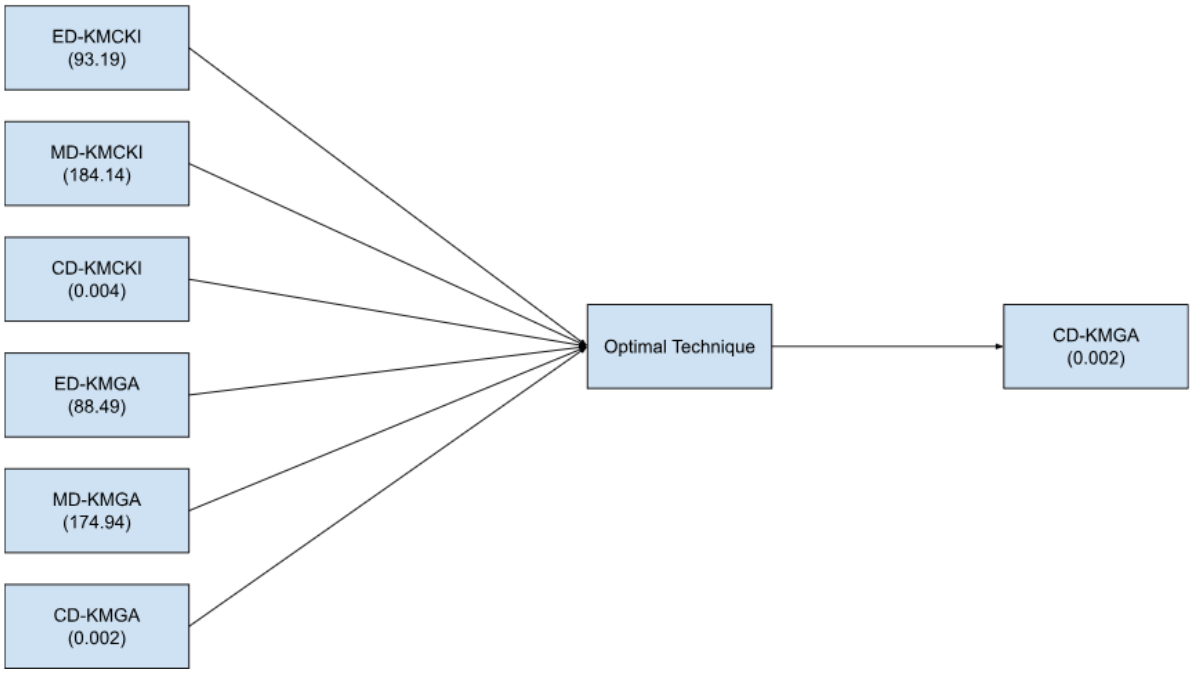


Figure 46. The Optimal Intelligent Technique to Determine the Best Centroid in the Clusters Regarding SE.

The study's findings indicate a discernible rise in electricity usage during January, February, November, and December. Furthermore, there was a notable reduction in energy consumption levels throughout June and July. Additionally, these results aid decision-makers in identifying the months during which energy consumption in public buildings experiences a notable increase. Therefore, the individuals residing in these structures are efficiently directed, as depicted in Figure 47.

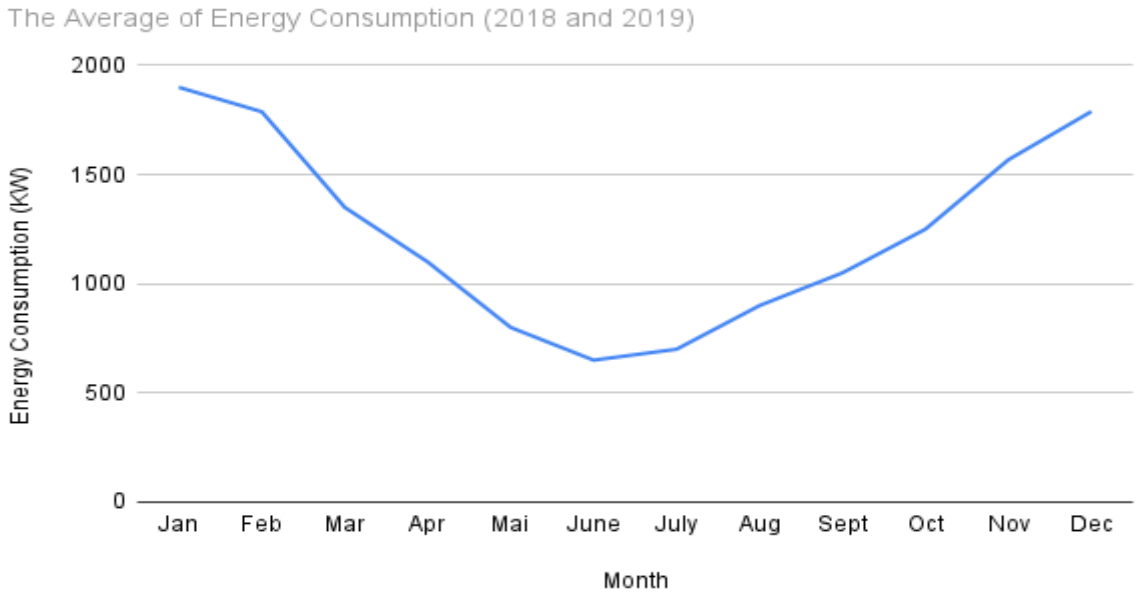


Figure 47. The Average of Energy consumption in (2018 and 2019)

Furthermore, it has been observed that several Municipalities had elevated energy consumption in specific months of 2018 and 2019, particularly in public buildings, as depicted in the figure 48.

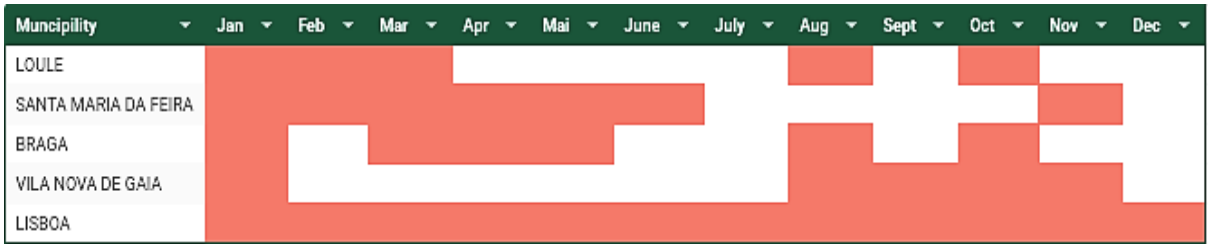


Figure 48. The Municipalities that Consumed High Energy Consumption in Public Buildings (2018 and 2019), Note: Red color(High energy), White color (low energy)

Finally, we compared our clustering results (KMGA) with state-of-the-art methods in previous works, where KMGA outperforms the state-of-the-art methods in terms of SE, as shown in figure 49.

5.1.3. Predictive Modeling with Deep Learning

We have created models employing deep learning methods, particularly CNN, to predict energy consumption levels by leveraging past data. The models above have exhibited considerable efficacy in predicting energy demand and facilitating enhanced resource allocation and administration.

We employed two deep learning models, namely CNN and CNN-GA. The CNN-GA model demonstrates superior performance compared to the CNN model in accurately estimating energy consumption levels of public buildings, as seen by higher accuracy and lower loss metrics. Furthermore, it can be observed from Figure 50 that CNN-GA exhibits superior performance in terms of MAE compared to the existing state-of-the-art methods.

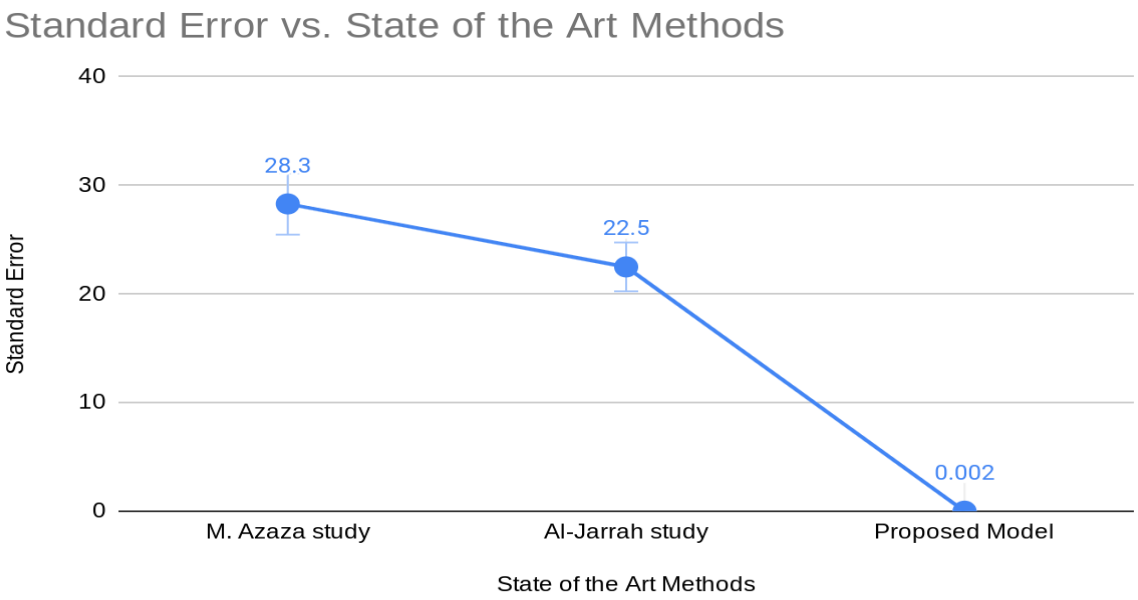


Figure 49. Comparison Between KMGA and State-of-the-Art Methods in terms of SE

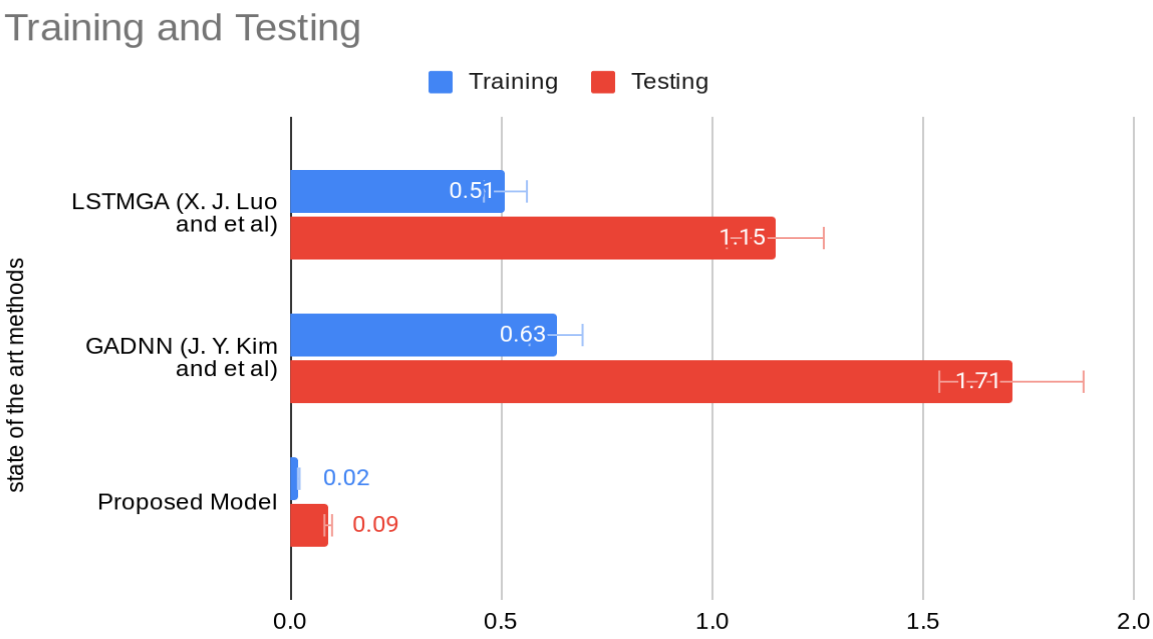


Figure 50. Comparison Between the Proposed Model (CNN-GA) and State-of-the-Art methods in terms of MAE

5.1.4. Identification of Optimal Settings

Through the examination of the correlation between various building factors, namely those outlined in Table 13, including active energy at specific time intervals (00h00-02h00; 06h00-08h00; 22h00-00h00), among others, and energy consumption patterns, we successfully determined the most advantageous configurations for promoting energy efficiency in operational practices. This information can guide building operators in making educated decisions to mitigate energy wastage. Furthermore, contracted electricity is a noteworthy aspect of Energy Consumption Data (ECD) since it provides

valuable insights into the varying energy consumption levels within public buildings across different periods.

By examining the clustering outcomes, several fundamental principles have been derived to aid stakeholders within the energy sector in Portugal in discerning the various typologies of public buildings, as depicted in Table 19. Energy consumption regulations assist decision-makers in identifying public buildings that require assistance for their occupants and facilitate the transition to alternative energy sources for such facilities.

If-then rules can help public buildings control their energy use in several ways, including:

1. The if-then rules are simple and understandable, rendering them accessible even to individuals lacking expertise in the subject matter. Therefore, both building managers and residents will experience enhanced comprehension of the impact of their decisions on energy use.

2. Subject to Modification, If-then rules can be adjusted to fulfill the specifications of a particular structure or organization. This enables the implementation of energy management strategies with a higher degree of specialization.

3. Real-time feedback can be provided to building managers and residents regarding their energy usage by implementing If-then rules. Individuals can modify their behavior and make informed decisions regarding energy expenditure due to acquiring enhanced knowledge.

4. Cost-effectiveness, the utilisation of if-then rules for energy management is considered a cost-effective strategy due to its compatibility with existing building management systems and sensors.

5. Energy conservation, If-then rules can mitigate energy consumption and reduce associated financial expenditures. Implementing real-time feedback mechanisms and promoting energy-efficient behavior can reduce energy consumption and financial savings for building management and occupants.

5.1.5. Potential for Energy Savings

The study revealed substantial energy conservation opportunities in buildings through implementing informed energy management systems utilising clustering and deep learning methodologies. The findings above underscore the need to employ sophisticated analytics techniques to optimize energy consumption and mitigate environmental consequences.

Accurate prediction of energy consumption is crucial for effective monthly building energy management. Precise estimation of peak and monthly demand is vital for successfully scheduling and managing energy devices, as it can enhance the building's energy utilization rate. With accurate building energy consumption predictions, building managers can improve their decision-making abilities in effectively managing various energy equipment.

Accurate energy usage predictions could be the basis for intelligent energy management and retrofitting initiatives to enhance building energy efficiency. Energy consumption predictions can provide valuable insights for building facility managers, enabling them to make informed decisions regarding future energy expenditures. This information can assist in evaluating the feasibility of transitioning to a more efficient pricing structure or implementing measures to reduce predicted energy usage.

5.2. Implications and Applications

The findings of our study have important implications for various stakeholders involved in energy consumption in buildings:

5.2.1. Building Operators

Our study offers valuable insights into optimizing energy consumption, resulting in financial savings and enhanced environmental sustainability. Using the recommended solutions, operators of buildings can attain enhanced sustainability in their operations and effectively mitigate their carbon emissions.

5.2.2. Energy Suppliers and Grid Operators

The study has produced predictive models that can assist energy suppliers and grid operators in anticipating fluctuations in demand and optimizing energy distribution accordingly. This phenomenon can potentially enhance energy efficiency and promote the development of a robust and resilient electrical grid system.

5.2.3. Policy Makers and Planners

The results of this study can contribute valuable insights to the formulation of energy-related laws and regulations. These insights can be crucial in promoting the widespread adoption of energy-efficient practices within the building sector. The field of urban planning can derive advantages from our research by incorporating energy consumption trends into the planning and construction of sustainable communities and structures.

5.3. Study Limitations

Despite the significance of our findings, there are some limitations to address in our study:

5.3.1. Data Availability and Quality

The validity and comprehensiveness of our results are contingent upon the accessibility and Caliber of the dataset employed. Subsequent investigations should gather complete and diverse datasets to enhance the analytical robustness.

5.3.2. Generalizability

The scope of our study was limited to a specific group of structures and a defined geographical region. To augment the generalizability of our findings, it is imperative to investigate a more comprehensive array of building typologies, climatic conditions, and geographical locations.

5.4. Discussion Summary

The study presented several noteworthy findings. Using clustering techniques facilitated the discernment of discrete energy consumption trends within structures. This enabled a more detailed comprehension of energy consumption and the possible determinants impacting it. Additionally, applying deep learning algorithms has shown considerable potential in accurately forecasting energy demand based on consumption patterns. The capacity to make accurate predictions might be advantageous for building operators to strategize and oversee energy resources efficiently. Furthermore, the study has successfully identified the most favorable configurations for achieving energy-efficient operations, presenting a promising opportunity for substantial reductions in energy consumption and the subsequent financial burdens.

The ramifications of the thesis findings extend to several stakeholders, encompassing building operators, energy providers, grid operators, policymakers, and planners. Using clustering and deep learning techniques can provide building operators with valuable insights, enabling them to make informed decisions to enhance energy efficiency within their facilities. Deep learning algorithms allow energy suppliers and grid operators to optimize their distribution efficiency. Policymakers and planners can utilize the research findings to inform the formulation of sustainable construction policies and the design of communities that are more energy efficient.

Chapter 6 – Conclusion and Future Work

6.1. Conclusion

We could summarize this section into three main phases, as follows:

In Chapter Two, this study conducted a thorough literature evaluation on the topic of categorizing and predicting the ECB with the goal of answering our five research questions. Before a more extensive manuscript analysis, text mining algorithms were used to determine the most used phrases in the energy and IC model domains. A bibliometric map was utilized to find the linkages between the most used terms in both domains. In our survey, we used a PRISMA technique to find 822 manuscripts and ended up analyzing 41. Our survey identified the most utilized IC models, particularly machine learning algorithms, used by the community to categorize and forecast the ECB. This research contributes to three areas. The first examines the factors that impact the ECB. The second offers a systematic overview of categorization and prediction strategies utilized in that setting. The final part addresses the evaluation criteria employed by these procedures.

In Chapter Three, this work proposed a novel hybrid intelligence model for identifying buildings' energy consumption levels (low, medium, and high), which was tested on an ECD in Portugal. We framed our investigation by posing four research questions that were thoroughly answered. A correlation coefficient analysis was done to determine the essential components (variables) that influence ECPB and comprehend the link between those factors to grasp our data better. An isolation forest was employed in the data preparation stage to remove outliers from the dataset. An interpolation method was also utilised to find compensation values or estimate unknown values based on related known values. In terms of our modeling approach, we first computed the number of energy consumption clusters in the dataset, and SOM, the Elbow method, and the Davis-Bouldin method all agreed on 3 as the figure for the found number of clusters (corresponding to low, medium, and high consumption).

Then we utilised KM using a GA to forecast each building's energy consumption cluster level. This work makes contributions in four areas. The first addresses elements that influence construction energy use. The second presents a novel approach for categorising energy use in public buildings into levels (for example, low, medium, and high). The third one examines the energy usage of public buildings in Portugal in 2018 and 2019 using real big data (77 996 public buildings in 238 Portuguese cities). As an example, we were able to identify localities with excessive energy use. We also determined monthly energy consumption patterns of buildings in 2018 and 2019. The final part extracts appropriate scientific If-Then principles to assist decisionmakers in rationalising and determining the most energy-consuming public buildings from a set of three values (low, medium, or high consumption).

In Chapter 4, This study provided a hybrid intelligent model to anticipate future energy consumption levels in public buildings, with an emphasis on meeting the four research objectives we set out to investigate. Raw data was collected monthly in 77 996 buildings spanning a wide range of public sectors and 238 cities in Portugal in 2018 and 2019. We used an isolation forest to remove values from our energy dataset that were outliers, and we interpolated to find compensation values or estimate unknown values based on related known values. Following data preparation, the final number of records used in this work was 1,222,695, corresponding to 26,624 public buildings, after excluding records of public lighting (because it is outside the scope of our study) and buildings that did not contain consumption data for the entire observed period of 24 months.

Then, each building's cluster labels were predicted using KM and a GA. Simultaneously, the Elbow technique and the Davis and Boulden approach were used to determine the optimal number of clusters. Furthermore, If-Then rules were created from KM data using a GA to assist in finding the buildings with the highest energy consumption. Finally, CNN and CNN-GA, two IC approaches, have been created to forecast future energy usage. This research has a significant impact in four areas. We provide a novel method for categorising the predicted ECPB (e.g., low, medium, and high). Our study is based on a huge data set of public building energy use collected in Portugal between 2018 and 2019. We derived good scientific If-Then principles for decision-makers to utilise in justifying public building energy use and identifying the largest energy hogs among them. Finally, we offer two IC models for forecasting future energy use, together with an assessment of the models' ACC and SER.

6.2. Future Work

Future research should focus on identifying essential factors influencing public building energy consumption using statistical methods such as multiple linear regression or logistic regression, as well as combining clustering and optimization techniques (grey wolf, lion, and whale optimization) to improve the ACC predictions for cluster labels describing building energy consumption (low - medium, and high).

As a result, there are still prospects for advancement in our research field. Other factors that affect energy usage in buildings (e.g., green roof, building envelope, internal and exterior factors) are suggested for future research. These parameters could be utilized to cluster and predict energy use. Our survey advises tackling ECB classification by combining clustering and optimization techniques, with the goal of categorizing the ECB into levels (low, medium, high). In terms of predicting the ECB, this study advises using machine learning methodologies from the deep learning family, such as transfer learning and deep forest, which are hot research trends.

By analyzing the results of our study, the following recommendations can be made in the future:

Expand the dataset: To enhance the reliability and generalizability of the findings, it is recommended to incorporate more extensive and more diverse datasets. This can include data from different building types and occupancy patterns. Increasing the dataset size will improve the accuracy of clustering techniques and the robustness of deep learning models.

Explore feature engineering: Consider incorporating more advanced feature engineering techniques to extract additional meaningful features from the energy consumption data. This can help in improving the accuracy and interpretability of the models. Features like weather data, occupancy patterns, and building characteristics (design architecture) can be considered to capture the underlying factors influencing energy consumption.

Evaluate real-time data: Incorporate real-time data collection and analysis techniques to capture dynamic changes in energy consumption patterns. This can be done by integrating sensor data and Internet of Things (IoT) devices in buildings. Real-time data enables the identification of immediate energy-saving opportunities and the timely adjustment of energy management strategies.

Consider socio-behavioral aspects: Energy consumption in buildings is influenced by human behavior. Consider incorporating socio-behavioral aspects in the analysis, such as occupancy schedules, user behaviours, and occupant preferences. This can help in designing targeted interventions and personalized energy-saving recommendations.

Long-term impact analysis: Extend the research to evaluate the long-term impact of energy-saving measures implemented based on clustering and deep learning techniques. This can provide insights into the effectiveness of the strategies and identify any potential issues or opportunities for improvement.

Collaboration with stakeholders: Engage with building operators, energy suppliers, and policymakers to ensure the practical implementation and adoption of the research findings. Collaborative efforts can lead to the development of effective energy management strategies, policies, and interventions at a larger scale.

By implementing these recommendations, further advancements can be achieved in utilizing clustering and deep learning techniques for energy consumption analysis in buildings. This will contribute to more energy-efficient operations, reduced energy costs, and a sustainable built environment.

References

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of Artificial Neural Network and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, *33*, 102–109. https://doi.org/10.1016/j.rser.2014.01.069

Aiello. (2018). DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. Energy and Buildings, 163(December), 58–69. https://doi.org/10.1016/j.enbuild.2017.12.040

Al-Wakeel, A., Wu, J., & Jenkins, N. (2017). K-Means Based Load Estimation of Domestic Smart Meter Measurements. Applied Energy, 194, 333–342. https://doi.org/10.1016/j.apenergy.2016.06.046

Azaza, M., & Wallin, F. (2017). Smart meter data clustering using consumption indicators: Responsibility factor and consumption variability. Energy Procedia, 142, 2236–2242. https://doi.org/10.1016/j.egypro.2017.12.624

Benachour, E., Draoui, B., Imine, B., & Asnoune, K. (2017). Numerical simulation of conjugate convection combined with the thermal conduction using a polynomial interpolation method. *Advances in Mechanical Engineering*, *9*(5), 1–7. https://doi.org/10.1177/1687814017700064

Cai, H., Shen, S., Lin, Q., Li, X., & Xiao, H. (2019). Predicting the Energy Consumption of Residential Buildings for Regional Electricity Supply-Side and Demand-Side Management. *IEEE Access*, *7*(July), 30386–30397. https://doi.org/10.1109/ACCESS.2019.2901257

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing, 307(May), 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

Ding, Z., Wang, Z., Hu, T., & Wang, H. (2022). A Comprehensive Study on Integrating Clustering with Regression for Short-Term Forecasting of Building Energy Consumption: Case Study of a Green Building. *Buildings*, *12*(10). https://doi.org/10.3390/buildings12101701

Ford, V., & Siraj, A. (2013). Clustering of smart meter data for disaggregation. *2013 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings*, (December 2013), 507–510. https://doi.org/10.1109/GlobalSIP.2013.6736926

Gouveia, J. P., & Seixas, J. (2016). Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy and Buildings*, *116*, 666–676. https://doi.org/10.1016/j.enbuild.2016.01.043

Hernández, L., Baladrón, C., Aguiar, J. M., Carro, B., & Sánchez-Esguevillas, A. (2012). Classification and clustering of electricity demand patterns in industrial parks. *Energies*, *5*(12), 5215–5228. https://doi.org/10.3390/en5125215

Kim, J., Naganathan, H., Moon, S.-Y., Chong, W. K. O., & Ariaratnam, S. T. (2017). Applications of Clustering and Isolation Forest Techniques in Real-Time Building Energy-Consumption Data: Application to LEED Certified Buildings. *Journal of Energy Engineering*, *143*(5), 04017052. https://doi.org/10.1061/(asce)ey.1943-7897.0000479

Lee, J., Kim, J., & Ko, W. (2019). Day-ahead electric load forecasting for the residential building with a small-size dataset based on a Self-Organizing Map and a stacking ensemble learning method. *Applied Sciences (Switzerland)*, *9*(6). https://doi.org/10.3390/app9061231

Liu, Q., & Ren, J. (2018). Research on technology clusters and the energy efficiency of energy-saving retrofits of existing office buildings in different climatic regions. *Energy, Sustainability and Society*, *8*(1). https://doi.org/10.1186/s13705-018-0165-0

Massana, J., Pous, C., Burgas, L., Melendez, J., & Colomer, J. (2016). Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. *Energy and Buildings*, *130*, 519–531. https://doi.org/10.1016/j.enbuild.2016.08.081

Miller, C., Nagy, Z., & Schlueter, A. (2018). A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews*, *81*(January), 1365–1377. https://doi.org/10.1016/j.rser.2017.05.124

Naji, S., Keivani, A., Shamshirband, S., Alengaram, U. J., Jumaat, M. Z., Mansor, Z., & Lee, M. (2016). Estimating building energy consumption using extreme learning machine method. *Energy*, *97*, 506–516. https://doi.org/10.1016/j.energy.2015.11.037

Nordahl, C., Boeva, V., Grahn, H., & Netz, M. P. (2019). Profiling of household residents' electricity consumption behavior using clustering analysis. *Lecture Notes in Computer Science (Including Subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *11540 LNCS*, 779–786. https://doi.org/10.1007/978-3-030-22750-0_78

Park, H. S., Lee, M., Kang, H., Hong, T., & Jeong, J. (2016). Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. *Applied Energy*, *173*, 225–237. https://doi.org/10.1016/j.apenergy.2016.04.035

Park, K. J., & Son, S. Y. (2019). A Novel Load Image Profile-Based Electricity Load Clustering Methodology. *IEEE Access*, *7*, 59048–59058. https://doi.org/10.1109/ACCESS.2019.2914216

Shi, G., Liu, D., & Wei, Q. (2016). Energy consumption prediction of office buildings based on echo state networks. *Neurocomputing*, *216*, 478–488. https://doi.org/10.1016/j.neucom.2016.08.004

Sternby, J., Thormarker, E., & Liljenstam, M. (2020). Anomaly detection forest. *Frontiers in Artificial Intelligence and Applications*, *325*, 1507–1514. https://doi.org/10.3233/FAIA200258

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, Vol. 13, pp. 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

Wen, L., Zhou, K., & Yang, S. (2019). A shape-based clustering method for pattern recognition of residential electricity consumption. *Journal of Cleaner Production*, *212*, 475–488. https://doi.org/10.1016/j.jclepro.2018.12.067

Zhao, D., Zhong, M., Zhang, X., & Su, X. (2016). Energy consumption predicting model of VRV (Variable refrigerant volume) system in office buildings based on data mining. *Energy*, *102*, 660–668. https://doi.org/10.1016/j.energy.2016.02.134

Abdelaziz, A., Santos, V., & Dias, M. S. (2021). Machine learning techniques in the energy consumption of buildings: A systematic literature review using text mining and bibliometric analysis. Energies, 14(22). https://doi.org/10.3390/en14227810

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, *33*, 102–109. https://doi.org/10.1016/j.rser.2014.01.069

Aiello. (2018). DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. *Energy and Buildings*, *163*(December), 58–69. https://doi.org/10.1016/j.enbuild.2017.12.040

Al-Jarrah, O. Y., Al-Hammadi, Y., Yoo, P. D., & Muhaidat, S. (2017). Multi-Layered Clustering for Power Consumption Profiling in Smart Grids. *IEEE Access*, *5*, 18459–18468. https://doi.org/10.1109/ACCESS.2017.2712258

Azaza, M., & Wallin, F. (2017). Smart meter data clustering using consumption indicators: Responsibility factor and consumption variability. *Energy Procedia*, *142*, 2236–2242. https://doi.org/10.1016/j.egypro.2017.12.624

Cai, H., Shen, S., Lin, Q., Li, X., & Xiao, H. (2019). Predicting the Energy Consumption of Residential Buildings for Regional Electricity Supply-Side and Demand-Side Management. *IEEE Access*, *7*(July), 30386–30397. https://doi.org/10.1109/ACCESS.2019.2901257

Chen, Z., Xiao, F., Guo, F., & Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy*, *9*(October 2022), 100123. https://doi.org/10.1016/j.adapen.2023.100123

de Santis, R. B., & Costa, M. A. (2020). Extended isolation forests for fault detection in small hydroelectric plants. *Sustainability (Switzerland)*, *12*(16). https://doi.org/10.3390/SU12166421

Ding, Z., Wang, Z., Hu, T., & Wang, H. (2022). A Comprehensive Study on Integrating Clustering with Regression for Short-Term Forecasting of Building Energy Consumption: Case Study of a Green Building. *Buildings*, *12*(10). https://doi.org/10.3390/buildings12101701

Dong, B., Yan, D., Li, Z., Jin, Y., Feng, X., & Fontenot, H. (2018). Modeling occupancy and behavior for better building design and operation—A critical review. *Building Simulation*, *11*(5), 899–921. https://doi.org/10.1007/s12273-018-0452-x

Ford, V., & Siraj, A. (2013). Clustering of smart meter data for disaggregation. *2013 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings*, (December 2013), 507–510. https://doi.org/10.1109/GlobalSIP.2013.6736926

Gouveia, J. P., & Seixas, J. (2016). Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy and Buildings*, *116*, 666–676. https://doi.org/10.1016/j.enbuild.2016.01.043

Granell, R., Axon, C. J., & Wallom, D. C. H. (2015). Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles. *IEEE Transactions on Power Systems*, *30*(6),

3217–3224. https://doi.org/10.1109/TPWRS.2014.2377213

Hariri, S., Kind, M. C., & Brunner, R. J. (2019). *Extended Isolation Forest with Randomly Oriented Hyperplanes*.

Hernández, L., Baladrón, C., Aguiar, J. M., Carro, B., & Sánchez-Esguevillas, A. (2012). Classification and clustering of electricity demand patterns in industrial parks. *Energies*, *5*(12), 5215–5228. https://doi.org/10.3390/en5125215

Hsu, D. (2015). Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Applied Energy*, *160*, 153–163. https://doi.org/10.1016/j.apenergy.2015.08.126

Ioannou, A. E., Kofinas, D., Spyropoulou, A., & Laspidou, C. (2017). Data mining for household water consumption analysis using SOM. *European Water*, *58*, 443–448.

Liu, Q., & Ren, J. (2018). Research on technology clusters and the energy efficiency of energy-saving retrofits of existing office buildings in different climatic regions. *Energy, Sustainability and Society*, *8*(1). https://doi.org/10.1186/s13705-018-0165-0

Naji, S., Keivani, A., Shamshirband, S., Alengaram, U. J., Jumaat, M. Z., Mansor, Z., & Lee, M. (2016). Estimating building energy consumption using extreme learning machine method. *Energy*, *97*, 506–516. https://doi.org/10.1016/j.energy.2015.11.037

Ouf, M. M., Gunay, H. B., & O'Brien, W. (2019). A method to generate design-sensitive occupant-related schedules for building performance simulations. *Science and Technology for the Built Environment*, *25*(2), 221–232. https://doi.org/10.1080/23744731.2018.1514855

Räsänen, T., Ruuskanen, J., & Kolehmainen, M. (2008). Reducing energy consumption by using SOMs to create more personalized electricity use information. *Applied Energy*, *85*(9), 830–840. https://doi.org/10.1016/j.apenergy.2007.10.012

Rhodes, J. D., Cole, W. J., Upshaw, C. R., Edgar, T. F., & Webber, M. E. (2014). Clustering analysis of residential electricity demand profiles. *Applied Energy*, *135*, 461–471. https://doi.org/10.1016/j.apenergy.2014.08.111

Swan, L. G., & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, Vol. 13, pp. 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

Zhao, H. X., & Magoulès, F. (2012). Feature selection for predicting building energy consumption based on statistical learning method. *Journal of Algorithms and Computational Technology*, *6*(1), 59–77. https://doi.org/10.1260/1748-3018.6.1.59

Arjunan, P., Poolla, K., & Miller, C. (2022). BEEM: Data-driven building energy benchmarking for Singapore. *Energy and Buildings*, *260*. https://doi.org/10.1016/j.enbuild.2022.111869

Bourhnane, S., Abid, M. R., Lghoul, R., Zine-Dine, K., Elkamoun, N., & Benhaddou, D. (2020). Machine learning for energy consumption prediction and scheduling in smart buildings. *SN Applied Sciences*, *2*(2), 1–10. https://doi.org/10.1007/s42452-020-2024-9

Cai, H., Shen, S., Lin, Q., Li, X., & Xiao, H. (2019). Predicting the Energy Consumption of Residential Buildings for Regional Electricity Supply-Side and Demand-Side Management. *IEEE Access*, *7*(July), 30386–30397. https://doi.org/10.1109/ACCESS.2019.2901257

Chaowen, H., & Dong, W. (2015). Prediction on hourly cooling load of buildings based on Neural Networks. *International Journal of Smart Home*, *9*(2), 35–52. https://doi.org/10.14257/ijsh.2015.9.2.04

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

Chen, Y., Chen, Z., Yuan, X., Su, L., & Li, K. (2022). Optimal Control Strategies for Demand Response in Buildings under Penetration of Renewable Energy. *Buildings*, *12*(3). https://doi.org/10.3390/buildings12030371

Chen, Y., Guo, M., Chen, Z., Chen, Z., & Ji, Y. (2022). Physical energy and data-driven models in building energy prediction: A review. *Energy Reports*, *8*, 2656–2671. https://doi.org/10.1016/j.egyr.2022.01.162

Chen, Z., Xiao, F., Guo, F., & Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy*, *9*(October 2022), 100123. https://doi.org/10.1016/j.adapen.2023.100123

Daim, T., Oliver, T., & Kim, J. (2013). Research and Technology Management in the Electricity Industry: Methods, Tools and Case Studies. *Green Energy and Technology*, *60*, 17–31.

https://doi.org/10.1007/978-1-4471-5097-8

Divina, F., Torres, J. F., García-Torres, M., Martínez-álvarez, F., & Troncoso, A. (2020). Hybridizing deep learning and neuroevolution: Application to the Spanish short-term electric energy consumption forecasting. *Applied Sciences (Switzerland)*, *10*(16). https://doi.org/10.3390/app10165487

Dong, Z., Liu, J., Liu, B., Li, K., & Li, X. (2021). Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification. *Energy and Buildings*, *241*, 110929. https://doi.org/10.1016/j.enbuild.2021.110929

Fathi, S., Srinivasan, R., Fenner, A., & Fathi, S. (2020). Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews*, *133*(August), 110287. https://doi.org/10.1016/j.rser.2020.110287

Fei, J., Chen, Y., Liu, L., & Fang, Y. (2022). Fuzzy Multiple Hidden Layer Recurrent Neural Control of Nonlinear System Using Terminal Sliding-Mode Controller. *IEEE Transactions on Cybernetics*, *52*(9), 9519–9534. https://doi.org/10.1109/TCYB.2021.3052234

Fu, Y., Li, Z., Zhang, H., & Xu, P. (2015). Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices. *Procedia Engineering*, *121*, 1016–1022. https://doi.org/10.1016/j.proeng.2015.09.097

Gouveia, J. P., Seixas, J., & Mestre, A. (2017). Daily electricity consumption profiles from smart meters - Proxies of behavior for space heating and cooling. *Energy*, *141*, 108–122. https://doi.org/10.1016/j.energy.2017.09.049

Helwig, N. E., Hong, S., & Hsiao-wecksler, E. T. (2020). Hybrid method for building energy consumption prediction based on limited data. IEEE PES/IAS PowerAfrica, 60–64.

Huang, S., Zuo, W., & Sohn, M. D. (2016). A Bayesian network model for predicting the cooling load of educational facilities. *ASHRAE and IBPSA-USA Building Simulation Conference*, *11*(1), 1–8.

Kim, T. Y., & Cho, S. B. (2019). Predicting residential energy consumption using CNN-LSTM NNs. *Energy*, *182*, 72–81. https://doi.org/10.1016/j.energy.2019.05.230

Li, Cheng, Yang, F., Xiao, Q., & Gao, Y. (2023). Climate Change and Its Impacts on Terrestrial Ecosystems: Recent Advances and Future Directions. *Atmosphere*, *14*(7), 10–12. https://doi.org/10.3390/atmos14071176

Li, Chengdong, Ding, Z., Zhao, D., Yi, J., & Zhang, G. (2017). Building energy consumption prediction: An extreme deep learning approach. *Energies*, *10*(10), 1–20. https://doi.org/10.3390/en10101525

Li, L., Sun, W., Hu, W., & Sun, Y. (2021). Impact of natural and social environmental factors on building energy consumption: Based on bibliometrics. *Journal of Building Engineering*, *37*(January). https://doi.org/10.1016/j.jobe.2020.102136

Luo, X. J., & Oyedele, L. O. (2021). Forecasting building energy consumption: Adaptive long-short term memory Neural Networks driven by genetic algorithm. *Advanced Engineering Informatics*, *50*(April), 101357. https://doi.org/10.1016/j.aei.2021.101357

Luo, X. J., Oyedele, L. O., Ajayi, A. O., Akinade, O. O., Owolabi, H. A., & Ahmed, A. (2020). Feature extraction and genetic algorithm enhanced adaptive deep Neural Network for energy consumption prediction in buildings. *Renewable and Sustainable Energy Reviews*, *131*(June), 109980. https://doi.org/10.1016/j.rser.2020.109980

McNeil, M. A., Karali, N., & Letschert, V. (2019). Forecasting Indonesia's electricity load through 2030 and peak demand reductions from appliance and lighting efficiency. *Energy for Sustainable Development*, *49*, 65–77. https://doi.org/10.1016/j.esd.2019.01.001

Müller, I. M. (2021). Feature selection for energy system modeling: Identification of relevant time series information. *Energy and AI*, *4*. https://doi.org/10.1016/j.egyai.2021.100057

Nada, S., & Hamed, M. (2019). Statistical Analysis for Economics of the Energy Development in North Zone of Cairo. *International Journal of Economics and Business Administration*, *5*(3), 140–160.

Nguyen, T. A., & Aiello, M. (2013). Energy intelligent buildings based on user activity : A survey. *Energy & Buildings*, *56*, 244–257. https://doi.org/10.1016/j.enbuild.2012.09.005

Nordahl, C., Boeva, V., Grahn, H., & Netz, M. P. (2019). Profiling of household residents' electricity consumption behavior using clustering analysis. *Lecture Notes in Computer Science (Including Subseries* Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11540 LNCS, 779–786. https://doi.org/10.1007/978-3-030-22750-0_78

Pham, A. D., Ngo, N. T., Ha Truong, T. T., Huynh, N. T., & Truong, N. S. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production*, *260*, 121082. https://doi.org/10.1016/j.jclepro.2020.121082

Qavidel Fard, Z., Zomorodian, Z. S., & Korsavi, S. S. (2022). Application of machine learning in thermal comfort studies: A review of methods, performance and challenges. *Energy and Buildings*, *256*. https://doi.org/10.1016/j.enbuild.2021.111771

Ruiz, L. G. B., Pegalajar, M. C., Arcucci, R., & Molina-Solana, M. (2020). A time-series clustering methodology for knowledge extraction in energy consumption data. *Expert Systems with Applications*, *160*, 113731. https://doi.org/10.1016/j.eswa.2020.113731

Runge, J., & Zmeureanu, R. (2019). Forecasting energy use in buildings using artificial Neural Networks: A review. *Energies*, *12*(17). https://doi.org/10.3390/en12173254

Salam, M. A., Yazdani, M. G., Wen, F., Rahman, Q. M., Malik, O. A., & Hasan, S. (2020). Modeling and Forecasting of Energy Demands for Household Applications. *Global Challenges*, *4*(1), 1900065. https://doi.org/10.1002/gch2.201900065

Science, A., Bhuiyan, M., & Image, M. (2022). *Vehicle Speed Prediction based on Road Status using Machine Learning Advanced Research in Energy and Engineering Vehicle Speed Prediction based on Road Status using Machine Learning*.

Serale, G., Fiorentini, M., & Noussan, M. (2020). Development of algorithms for building energy efficiency. *Start-Up Creation: The Smart Eco-Efficient Built Environment, Second Edition*, 267–290. https://doi.org/10.1016/B978-0-12-819946-6.00011-4

Wu, H., Huang, A., & Sutherland, J. W. (2022). Layer-wise relevance propagation for interpreting LSTM-RNN decisions in predictive maintenance. International Journal of Advanced Manufacturing Technology, 118(3–4), 963–978. https://doi.org/10.1007/s00170-021-07911-9

Banihashemi, S., Ding, G., & Wang, J. (2017). Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption. Energy Procedia, 110(December 2016), 371–376. https://doi.org/10.1016/j.egypro.2017.03.155

Berardi, U. (2015). Building Energy Consumption in US, EU, and BRIC Countries. Procedia Engineering, 118, 128–136. https://doi.org/10.1016/j.proeng.2015.08.411

Berriel, R. F., Lopes, A. T., Rodrigues, A., Varejao, F. M., & Oliveira-Santos, T. (2017). Monthly energy consumption forecast: A deep learning approach. Proceedings of the International Joint Conference on Neural Networks, 2017-May, 4283–4290. https://doi.org/10.1109/IJCNN.2017.7966398

Bhattacharjee, S., & Reichard, G. (2011). Socio-economic factors affecting individual household energy consumption: A systematic review. ASME 2011 5th International Conference on Energy Sustainability, ES 2011, (PARTS A, B, AND C), 891–901. https://doi.org/10.1115/ES2011-54615

Bogner, K., Pappenberger, F., & Zappa, M. (2019). Machine Learning Techniques for Predicting the Energy Consumption/Production and Its Uncertainties Driven by Meteorological Observations and Forecasts. Sustainability, 11(12), 3328. https://doi.org/10.3390/su11123328

Carbonare, N., Pflug, T., & Wagner, A. (2018). Clustering the occupant behavior in residential buildings: A method comparison. *Bauphysik*, *40*(6), 427–433. https://doi.org/10.1002/bapi.201800023

E. Agência. (2018). Energy efficiency trends and policies in Portugal. Agência para a Energia, 1(1), 234–251.

Delzendeh, E., Wu, S., Lee, A., & Zhou, Y. (2017). The impact of occupants' behaviours on building energy analysis: A research review. *Renewable and Sustainable Energy Reviews*, *80*(May), 1061–1071. https://doi.org/10.1016/j.rser.2017.05.264

Diao, L., Sun, Y., Chen, Z., & Chen, J. (2017). Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy and Buildings*, *147*, 47–66. https://doi.org/10.1016/j.enbuild.2017.04.072

Djenouri, D., Laidi, R., Djenouri, Y., & Balasingham, I. (2019). Machine learning for smart building applications: Review and taxonomy. *ACM Computing Surveys*, *52*(2). https://doi.org/10.1145/3311950

Kleszcz-Szczyrba, R. (2010). "Pomagać sobą" - Rozważania na temat czynników niespecyficznych w psychoterapii związanych z osobą psychoterapeuty. *Psychoterapia*, Vol. 1, pp. 61–72.

Zhang, M., & Bai, C. (2018). Exploring the in fl uencing factors and decoupling state of residential energy consumption in Shandong. *Journal of Cleaner Production*, *194*, 253–262. https://doi.org/10.1016/j.jclepro.2018.05.122

Javaid, N., Ullah, I., Akbar, M., Iqbal, Z., Khan, F. A., Alrajeh, N., & Alabed, M. S. (2017). An Intelligent Load Management System with Renewable Energy Integration for Smart Homes. *IEEE Access*, *5*(c), 13587–13600. https://doi.org/10.1109/ACCESS.2017.2715225

Jozi, A., Pinto, T., Praça, I., & Vale, Z. (2019). Decision support application for energy consumption forecasting. *Applied Sciences (Switzerland)*, *9*(4). https://doi.org/10.3390/app9040699

Li, W. T., Yuen, C., Ul Hassan, N., Tushar, W., Wen, C. K., Wood, K. L., … Liu, X. (2015). Demand response management for residential smart grid: From theory to practice. *IEEE Access*, *3*, 2431–2440. https://doi.org/10.1109/ACCESS.2015.2503379

Qamar, M., & Khosravi, A. (2015). A review on arti fi cial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews*, *50*, 1352–1372. https://doi.org/10.1016/j.rser.2015.04.065

Bourdeau, M., Zhai, X. qiang, Nefzaoui, E., Guo, X., & Chatellier, P. (2019). Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society*, *48*(February), 101533. https://doi.org/10.1016/j.scs.2019.101533

Qolomany, B., Al-Fuqaha, A., Gupta, A., Benhaddou, D., Alwajidi, S., Qadir, J., & Fong, A. C. (2019). Leveraging Machine Learning and Big Data for Smart Buildings: A Comprehensive Survey. *IEEE Access*, *7*, 90316–90356. https://doi.org/10.1109/ACCESS.2019.2926642

Vázquez-Canteli, J. R., & Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*, *235*(April 2018), 1072–1089. https://doi.org/10.1016/j.apenergy.2018.11.002

Mason, K., & Grijalva, S. (2019). A review of reinforcement learning for autonomous building energy management. *Computers and Electrical Engineering*, *78*, 300–312. https://doi.org/10.1016/j.compeleceng.2019.07.019

Guyot, D., Giraud, F., Simon, F., Corgier, D., Marvillet, C., & Tremeac, B. (2019). Overview of the use of artificial Neural Networks for energy-related applications in the building sector. *International Journal of Energy Research*, *43*(13), 6680–6720. https://doi.org/10.1002/er.4706

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*(January 2016), 1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Mosavi, A., & Bahmani, A. (2019). Energy consumption prediction using m[1] A. Mosavi and A. Bahmani, "Energy consumption prediction using machine learning; a review," Energies, no. March, pp. 1–63, 2019.achine learning; a review. *Energies*, (March), 1–63. https://doi.org/10.20944/preprints201903.0131.v1

Perera, A. T. D., & Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, *137*(December 2020), 110618. https://doi.org/10.1016/j.rser.2020.110618

Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., & Gandomi, A. H. (2021). The Arithmetic Optimization Algorithm. *Computer Methods in Applied Mechanics and Engineering*, *376*, 113609. https://doi.org/10.1016/j.cma.2020.113609

Kacprzyk, J. (2014). Studies in computational intelligence. *Studies in Computational Intelligence*, *534*, 1–292. https://doi.org/10.1007/978-3-319-03419-5

Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hussien, A. G., Khasawneh, A. M., Alshinwan, M., & Houssein, E. H. (2020). Nature-inspired optimization algorithms for text document clustering—a comprehensive analysis. *Algorithms*, *13*(12), 1–32. https://doi.org/10.3390/a13120345

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, Vol. 6. https://doi.org/10.1371/journal.pmed.1000097

Trianni, A., Merigó, J. M., & Bertoldi, P. (2018). *Ten years of Energy Efficiency : a bibliometric analysis*. 1917–1939.

Ma, X., Wang, M., & Li, C. (2020). A summary on research of household energy consumption: A bibliometric analysis. *Sustainability (Switzerland)*, *12*(1). https://doi.org/10.3390/su12010316

Zhang, W. Z., Elgendy, a I. A., Hammad, M., Iliyasu, A. M., Du, X., Guizani, M., & El-Latif, A. A. A. (2020). Secure and Optimized Load Balancing for Multi-Tier IoT and Edge-Cloud Computing Systems. *IEEE Internet of Things Journal*, *4662*(c), 1–14. https://doi.org/10.1109/JIOT.2020.3042433

Wu, Y., Liu, Y., Ahmed, S. H., Peng, J., & El-Latif, A. A. A. (2020). Dominant Data Set Selection Algorithms for Electricity Consumption Time-Series Data Analysis Based on Affine Transformation. *IEEE Internet of Things Journal*, *7*(5), 4347–4360. https://doi.org/10.1109/JIOT.2019.2946753

Ahmed A.Abd El-Latif, BassemAbd-El-Atty, IrfanMehmood, KhanMuhammad, Salvador E.Venegas-Andraca, JialiangPeng (2021). Quantum-Inspired Blockchain-Based Cybersecurity: Securing Smart Edge Utilities in IoT-Based Smart Cities. Information Processing & Management, Volume 58, Issue 4, 102549. https://doi.org/10.1016/j.ipm.2021.102549

Andrew .K (2023). Exploratory Bibliometrics: Using VOSviewer as a Preliminary Research Tool. Publications, Volume 11, Issue 1. https://doi.org/10.3390/publications11010010

Guerra Santin, O. (2013). Occupant behaviour in energy efficient dwellings: Evidence of a rebound effect. *Journal of Housing and the Built Environment*, *28*(2), 311–327. https://doi.org/10.1007/s10901-012-9297-2

Mancini, F., Basso, G. Lo, & De Santoli, L. (2019). Energy use in residential buildings: Characterisation for identifying flexible loads by means of a questionnaire survey. *Energies*, *12*(11). https://doi.org/10.3390/en12112055

Csoknyai, T., Legardeur, J., Akle, A. A., & Horváth, M. (2019). Analysis of energy consumption profiles in residential buildings and impact assessment of a serious game on occupants' behavior. *Energy and Buildings*, *196*, 1–20. https://doi.org/10.1016/j.enbuild.2019.05.009

Mardookhy, M., Sawhney, R., Ji, S., Zhu, X., & Zhou, W. (2014). A study of energy efficiency in residential buildings in Knoxville, Tennessee. *Journal of Cleaner Production*, *85*, 241–249. https://doi.org/10.1016/j.jclepro.2013.09.025

Cao, X., Dai, X., & Liu, J. (2016). Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. *Energy and Buildings*, *128*, 198–213. https://doi.org/10.1016/j.enbuild.2016.06.089

Chang, C., Zhu, N., Yang, K., & Yang, F. (2018). Data and analytics for heating energy consumption of residential buildings: The case of a severe cold climate region of China. *Energy and Buildings*, *172*, 104–115. https://doi.org/10.1016/j.enbuild.2018.04.037

Hannan, M. A., Faisal, M., Ker, P. J., Mun, L. H., Parvin, K., Mahlia, T. M. I., & Blaabjerg, F. (2018). A review of internet of energy based building energy management systems: Issues and recommendations. *IEEE Access*, *6*(c), 38997–39014. https://doi.org/10.1109/ACCESS.2018.2852811

Nepal, B., Yamaha, M., Sahashi, H., & Yokoe, A. (2019). Analysis of building electricity use pattern using K-Means Clustering algorithm by determination of better initial centroids and number of clusters. *Energies*, *12*(12). https://doi.org/10.3390/en12122451

Pan, S., Wang, X., Wei, Y., Zhang, X., Gal, C., Ren, G., … Liu, J. (2017). Cluster analysis for occupant-behavior based electricity load patterns in buildings: A case study in Shanghai residences. *Building Simulation*, *10*(6), 889–898. https://doi.org/10.1007/s12273-017-0377-9

Tureczek, A., Nielsen, P. S., & Madsen, H. (2018). Electricity consumption clustering using smart meter data. *Energies*, *11*(4). https://doi.org/10.3390/en11040859

Wahid, F., Ghazali, R., Shah, A. S., & Fayaz, M. (2017). Prediction of Energy Consumption in the Buildings Using Multi-Layer Perceptron and Random Forest. *International Journal of Advanced Science and Technology*, *101*, 13–22. https://doi.org/10.14257/ijast.2017.101.02

Moretti, E., Nassuato, L., & Bordoni, G. (2019). Development of Regression Models to Predict Energy Consumption in Industrial Sites: The Case Study of a Manufacturing Company in the Central Italy. *TECNICA ITALIANA-Italian Journal of Engineering Science*, *63*(2–4), 343–348. https://doi.org/10.18280/ti-ijes.632-433

Edwards, R. E., New, J., & Parker, L. E. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, *49*, 591–603. https://doi.org/10.1016/j.enbuild.2012.03.010

Zekić-Sušac, M., Mitrović, S., & Has, A. (2020). Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International Journal of Information Management*, (January), 102074. https://doi.org/10.1016/j.ijinfomgt.2020.102074

Kim, S., Jung, S., & Baek, S. M. (2019). A model for predicting energy usage pattern types with energy consumption information according to the behaviors of single-person households in South Korea. *Sustainability (Switzerland)*, *11*(1). https://doi.org/10.3390/su11010245

Aqlan, F., Ahmed, A., Srihari, K., & Khasawneh, M. T. (2014). Integrating artificial Neural Networks and cluster analysis to assess energy efficiency of buildings. *IIE Annual Conference and Expo 2014*, (November 2015), 3936–3943.

Jovanović, R., & Sretenović, A. A. (2017). Ensemble of radial basis Neural Networks with K-Means Clustering for heating energy consumption prediction. *FME Transactions*, *45*(1), 51–57. https://doi.org/10.5937/fmet1701051J

Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. *Visualization in Engineering*, *6*(1). https://doi.org/10.1186/s40327-018-0064-7

Tang, W. J., Lee, X. L., Wang, H., & Yang, H. T. (2019). Leveraging Socioeconomic Information and Deep Learning for Residential Load Pattern Prediction. *Proceedings of 2019 IEEE PES Innovative Smart Grid Technologies Europe, ISGT-Europe 2019*. https://doi.org/10.1109/ISGTEurope.2019.8905483

Gajowniczek, K., & Zabkowski, T. (2018). Simulation Study on Clustering Approaches for Short-Term Electricity Forecasting. *Complexity*, *2018*(iii). https://doi.org/10.1155/2018/3683969