

**ENHANCING ENERGY EFFICIENCY IN RESIDENTIAL BUILDINGS
AT THE DESIGN STAGE THROUGH STATISTICAL AND
MACHINE LEARNING MODELS**

By

RAZAK A. OLU-AJAYI

**A thesis submitted to the University of Hertfordshire in partial
fulfilment of the requirements for the degree of
Doctor of Philosophy**

June 2024

ABSTRACT

The high proportion of energy consumed in buildings has led to significant environmental problems that negatively impact human existence. It is noted that the construction of energy-efficient buildings can help reduce the overall energy consumed in buildings. The prediction of building energy use is largely proclaimed to be a method for energy conservation and improved decision-making towards decreasing energy usage. Statistical and Machine Learning (ML) methods are recognised as highly effective for producing desired outcomes in prediction tasks. Consequently, ML has been extensively applied in studies focusing on the energy consumption of operational buildings. However, few studies explore the suitability of ML algorithms for predicting potential building energy consumption during the early design stage to facilitate the construction of more energy-efficient buildings.

This research developed a back-to-front model for building designers, using statistical and machine learning algorithms. Embracing a positivist paradigm due to its objective stance, allows for a rigorous investigation and experimental analysis of hypotheses and objective evaluation of various models. This research includes evaluating different feature selection impacts on models for classification and regression tasks, assessing various statistical and AI tools across several criteria within the building energy research domain, and comprehensively reviewing studies on various factors influencing energy use in buildings, among other investigations and analysis.

A key finding is that Gradient Boosting (GB) is identified as the most effective model in terms of both accuracy and computational efficiency. Through extensive investigation and analysis, GB emerged as the optimal choice for building an energy prediction model.

A significant contribution of this research is the development of a back-to-front model that allows building designers to specify target energy consumption values. By inputting values for relevant parameters into the optimization model, designers can obtain optimal values or specifications for building features required to achieve the desired energy consumption outcomes. Remarkably, the model produces results in less than five minutes. This will essentially revolutionize energy assessment at the conceptual stage of building development sustainably.

This research not only advances the theoretical understanding of building energy consumption prediction but also engenders practical tools for architects, engineers, and stakeholders to develop a more sustainable and efficient building. The integration of such a model using statistical and machine learning approaches into the design stage marks a significant step towards attaining environmentally friendly and economically viable buildings.

DEDICATION

With utmost reverence and gratitude, this thesis is firstly dedicated to Almighty Allah, the Most Merciful and Compassionate, whose infinite blessings and guidance have illuminated my path and sustained me throughout this journey. Subsequently, to my beloved parents, Alhaji and Alhaja R.A Olu-Ajayi, your prayers, guidance, love, sacrifices and unwavering faith in me have been a beacon of strength and inspiration and I am forever grateful for your endless love and guidance.

To my amazing wife, Mololuwa Temitope Olu-Ajayi, whose patience, understanding, and unwavering support have been a source of immense comfort and motivation. Thank you for always standing by my side and believing in me.

To my lovely sisters, Mariam Olu-Ajayi, Zainab Olu-Ajayi, Bisoye Sonaike, and Ibironke Solabi, your constant encouragement and belief in my abilities have been a source of strength. Thank you for always cheering me on and being there for me, no matter the challenges I faced.

This thesis is dedicated to each and every one of you, with heartfelt gratitude and appreciation. Thank you for being part of this incredible journey.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who has contributed to the completion of this doctoral thesis.

First and foremost, I am deeply indebted to my supervisor, Prof, Hafiz Alaka whose expertise, guidance, and unwavering support have been instrumental throughout this journey. His insightful feedback and encouragement have helped shape this work into its final form. I cannot thank him enough for all his efforts.

I am grateful to my second supervisor Dr Ketty Grishikashvili, for her valuable feedback, constructive criticism, and scholarly insights that have enriched the quality of this thesis. I also acknowledge the Doctoral Research Tutor Dr Francesca Gagliardi for her support throughout my research.

I am thankful to my colleagues (Habeeb Balogun, Christian Egwim, Ismail Sulaimon, Wasiu Yusuf, Godoyon Wusu, Muideen Adegoke and fellow researchers for their support, engaging discussions, and collaborative spirit, which have created an intellectually stimulating environment.

My deepest appreciation goes to my family for their unwavering love, encouragement, and understanding throughout this journey. Your support has been my source of strength and motivation. To all those who have contributed in ways whether big or small, I am truly grateful.

TABLE OF CONTENTS

ABSTRACT.....	II
DEDICATION	IV
ACKNOWLEDGEMENT	V
LIST OF TABLES	XIV
LIST OF FIGURES.....	XVI
ABBREVIATION	XIX
CHAPTER ONE	1
1.0 INTRODUCTION	1
1.1 CHAPTER OVERVIEW	1
1.2 BACKGROUND.....	1
1.3 RESEARCH PROBLEM AND GAP IN KNOWLEDGE.....	3
1.4 RESEARCH AIM AND OBJECTIVES.....	5
<i>1.4.1 Research Questions.....</i>	6
<i>1.4.2 Research Hypothesis</i>	6
<i>1.4.3 Research Methodology</i>	14
1.5 RESEARCH FRAMEWORK	15
1.6 RESEARCH CONTRIBUTIONS	17
1.7 JUSTIFICATION OF STUDIES.....	18
1.8 RESEARCH SCOPE AND LIMITATIONS	18
1.9 THESIS STRUCTURE	18
1.10 CHAPTER SUMMARY	22
CHAPTER TWO	23
2.0 OVERVIEW OF BUILDING ENERGY CONSUMPTION PREDICTION: METHODS, SIMULATIONS, TECHNIQUES, THEORY AND PERFORMANCE FACTORS	23
2.1 CHAPTER OVERVIEW	23

2.2 BUILDING ENERGY CONSUMPTION	23
2.3 ENERGY SIMULATION TOOLS (PHYSICAL METHOD)	26
2.4 ARTIFICIAL INTELLIGENCE-BASED TOOLS (DATA DRIVEN METHOD).....	28
2.4.1 <i>Machine Learning Algorithms</i>	30
2.4.1.1 <i>Types of Machine Learning Algorithms</i>	30
2.4.1.2 <i>Quantity of the Data</i>	36
2.4.1.3 <i>Quality of the Data</i>	37
2.5 THEORY REVIEW	38
2.5.1 <i>Forecasting Theory</i>	38
2.6 CHAPTER SUMMARY	39
CHAPTER THREE	40
3.0 SYSTEMATIC LITERATURE REVIEW AND THEORITICAL FRAMEWORK: STATISTICAL AND ARTIFICIAL INTELLIGENCE BASED TOOLS FOR BUILDING ENERGY CONSUMPTION PREDICTION.....	40
3.1 CHAPTER OVERVIEW	40
3.2 SIGNIFICANCE OF TOOL EVALUATION	40
3.3 EXISTING REVIEW STUDIES ON BUILDING ENERGY PREDICTION TOOLS	43
3.4 METHODOLOGY	44
3.4.1 <i>Data Collection</i>	46
3.4.1.1 <i>Search for Research Articles</i>	47
3.4.1.2 <i>Inclusion and Exclusion Criteria</i>	48
3.5 BIBLIOMETRIC ANALYSIS.....	49
3.5.1 <i>Publication Trends Analysis</i>	49
3.5.2 <i>Keywords Co-Occurrence Analysis</i>	50
3.5.3 <i>Top Journal Analysis</i>	52
3.5.4 <i>Global Collaboration Analysis</i>	54
3.6 SYSTEMATIC ANALYSIS	56
3.7 STATISTICAL AND AI-BASED TOOLS.....	56
3.7.1 <i>Relevant Criteria</i>	59
3.8 RESULT AND DISCUSSION	60

3.8.1 Error Rate.....	69
3.8.2 Building Type.....	75
3.8.3 Energy Type.....	77
3.8.5 Temporal Granularities	78
3.8.6 Data Size	79
3.8.7 Feature Selection.....	80
3.8.8 Proposed Framework	80
3.8.9 Theoretical and Practical Implications	83
3.9 CHAPTER SUMMARY	84
CHAPTER FOUR	86
4.0 SYSTEMATIC LITERATURE REVIEW: ANALYSIS OF FEATURES INFLUENCING BUILDING ENERGY PERFORMANCE.....	86
4.1 CHAPTER OVERVIEW	86
4.2 SIGNIFICANCE OF FEATURE ANALYSIS	86
4.3 METHODOLOGY	89
4.3.1 <i>Data Collection</i>	91
4.4 BIBLIOMETRIC ANALYSIS.....	93
4.4.1 <i>Publication Trends Analysis</i>	94
4.4.2 <i>Keywords Co-Occurrence Analysis</i>	95
4.4.3 <i>Global Collaboration Analysis</i>	96
4.5 THE FEATURES	98
4.6 RESULT AND DISCUSSION	100
4.6.1 <i>Impact of Building Floor</i>	107
4.6.2 <i>Impact of Window</i>	108
4.6.3 <i>Impact of Roof</i>	109
4.6.4 <i>Impact of Wall</i>	110
4.6.5 <i>Impact of Weather</i>	111
4.6.6 <i>Impact of Building Orientation</i>	112
4.6.7 <i>Impact of Occupancy</i>	113
4.6 CHAPTER SUMMARY	113
CHAPTER FIVE	114

5.0 RESEARCH METHODOLOGY	114
5.1 CHAPTER OVERVIEW	114
5.2 RESEARCH PHILOSOPHY AND APPROACH	114
5.2.1 <i>Positivism and Interpretivism</i>	115
5.2.2 <i>Ontological and Epistemological assumptions</i>	116
5.3 RESEARCH METHOD	116
5.4 RESEARCH APPROACH	119
5.5 DATA COLLECTION	120
5.5.1 <i>Building Metadata</i>	122
5.5.2 <i>Meteorological Data</i>	124
5.5.3 <i>Energy Data</i>	125
5.6 CHAPTER SUMMARY	128
CHAPTER SIX	129
6.0 FEATURE SELECTION EVALUATION FOR BUILDING ENERGY CONSUMPTION PREDICTION	129
6.1 CHAPTER OVERVIEW	129
6.2 SIGNIFICANCE OF FEATURE SELECTION	129
6.3 METHODOLOGY FOR FEATURE SELECTION.....	133
6.4 FEATURE SELECTION METHODS	135
6.4.1 <i>Filters</i>	135
6.4.2 <i>Wrappers</i>	135
6.4.3 <i>Embedded</i>	136
6.5 RESULT AND DISCUSSION	136
6.5.1 <i>Classification</i>	137
6.5.1.1 Feature Selection Analysis	137
6.5.1.2 Correlation Analysis (Classification).....	141
6.5.2 <i>Regression</i>	142
6.5.2.1 Feature Selection Analysis	142
6.5.2.2 Correlation Analysis	146

<i>6.5.5 H0: Openings (such as windows and doors) have the most Significant Effect on Building Energy Consumption.</i>	147
6.8 CHAPTER SUMMARY	151
CHAPTER SEVEN.....	152
7.0 DEVELOPMENT OF AI/ML BUILDING ENERGY CONSUMPTION PREDICTION MODEL...	152
7.1 CHAPTER OVERVIEW	152
7.2 BACKGROUND ON ENERGY PREDICTION MODEL.....	152
7.3 STATISTICAL AND MACHINE LEARNING TOOLS.....	154
7.4 DATA PRE-PROCESSING	157
7.4.1 Data Merging	158
7.4.2 Data Cleaning	158
7.4.3 Data Conversion	158
7.4.4 Data Normalization	158
7.5 MODEL DEVELOPMENT.....	159
7.5.1 Handling Imbalanced Datasets with SMOTE.....	161
7.6.2 Overview of Classification and Regression Models.....	162
7.6.2.1 Classification Model.....	162
7.6.2.2 Regression Model	164
7.7. MODEL EVALUATION.....	164
7.7.1 Evaluation of Classification Models	165
7.7.2 Evaluation of Regression Models	166
7.8 RESULT AND DISCUSSION	167
7.8.1 Feature Selection Impact on Model Performance	167
7.8.1.1 Regression Model Performance with and without Feature Selection Methods.	168
7.8.1.2 Classification Model Performance with and without Feature Selection Methods.	169
7.8.2 H1: Feature Selection does not influence Regression Model Performance	170
7.8.3 H2: Feature Selection has Positive Impacts on Machine Learning Energy Prediction Model Performance.	175

<i>7.8.4 H3: ML Feature Selection Methods lead to Better Performance of Machine Learning Prediction Models than Statistical Feature Selection Methods.....</i>	176
<i>7.8.5 H4: Using the Same Algorithm for Feature Selection and Prediction leads to better Model Performance than Using Different Algorithms.</i>	176
<i>7.8.6 H5: Weather Data Does not Significantly Improve ML Model Performance for Building Energy Consumption Prediction.</i>	179
<i>7.8.7 Key Feature Impact Analysis</i>	181
<i>7.8.7.1 Impact on Classification Model Performance</i>	182
<i>7.8.7.2 Impact on Regression Model Performance</i>	189
<i>7.8.8 H6: Machine Learning Algorithms produce better performing Building Energy Prediction Performance Models than Statistical Method.</i>	195
<i>7.8.9 H7: Deep Learning Algorithm outperforms Classical Machine Learning Algorithms.</i>	
<i>196</i>	
<i>7.8.10 Reliability Analysis (Data Size)</i>	197
<i>7.8.10.1 Data Size Impact on Classification Model Performance.....</i>	199
<i>7.8.10.2 Data Size Impact on Regression Model Performance</i>	203
<i>7.8.11 H8: Larger Data Size only Improve the Model Performance for certain ML Algorithms.....</i>	206
<i>7.8.12 Big Data: Model Performance Analysis</i>	208
<i>7.8.13 H9: Statistical/ML Tools can Assess Energy Consumption Faster than the Traditional Method</i>	210
<i>7.8.14 Hyperparameter Tuning for Model Optimization</i>	211
<i>7.8.14.1 Classification Model Optimization – Hyperparameter Tunning</i>	211
<i>7.8.14.2 Regression Model Optimization – Hyperparameter Tunning.....</i>	219
<i>7.9 CHAPTER IMPLICATIONS AND KEY INSIGHTS.....</i>	221
<i>7.9.1 Theoretical Implications.....</i>	222
<i>7.9.2 Practical Implications</i>	223
<i>7.10 CHAPTER SUMMARY</i>	224
CHAPTER EIGHT	225

8.0 REVERSE ENGINEERED SYSTEM AND VALIDATED FRAMEWORK FOR BUILDING ENERGY PREDICTION.....	225
8.1 CHAPTER OVERVIEW	225
8.2 VALIDATED FRAMEWORK	225
8.3 ENERGY OPTIMIZER.....	229
8.3.1 <i>Implementation Strategy</i>	230
8.3.2 <i>Optimization Approach</i>	232
8.3.3 <i>Energy Consumption Prediction Model Formula</i>	233
8.4 REVERSE-ENGINEERED SYSTEM.....	234
8.5 CHAPTER SUMMARY	237
CHAPTER NINE	238
9.0 CONCLUSIONS AND RECOMMENDATIONS.....	238
9.1 CHAPTER OVERVIEW	238
9.2 REVIEW OF RESEARCH OBJECTIVES AND CONCLUSIONS	238
9.2.1 <i>Objective One: To Establish the Key Features that Influence Energy Consumption In Buildings Using a Systematic Literature Review.</i>	239
9.2.2 <i>Objective Two: To Establish and Validate the Most Prominent Statistical And AI/ML Algorithms for Building Energy Consumption Prediction Using a Systematic Literature Review.</i>	239
9.2.3 <i>Objective Three: To Establish the Minimum Data Size Required for Developing an Efficient AI/ML Energy Prediction Model, by Benchmarking the Performance of ML Algorithms with Varying Data Sizes</i>	240
9.2.4 <i>Objective Four: To Identify the Most Accurate and Efficient Building Energy Prediction Model and The Best Performing Transparent Statistical and AI/ML Models</i> 241	
9.2.5 <i>Objective Five: To Develop an Optimization Model Using the Best Performing Transparent Model From Objective 4.</i>	242
9.3 CONTRIBUTIONS AND SIGNIFICANCE OF THE RESEARCH	242
9.3.1 <i>Theoretical Implications of Research</i>	242
9.3.2 <i>Practical Implications of Research</i>	245
9.4 LIMITATIONS OF THE RESEARCH	247
9.5 RECOMMENDATIONS FOR FUTURE RESEARCH	248

9.6 CHAPTER SUMMARY	249
REFERENCES.....	250
APPENDIX A: LIST OF RESEARCH PUBLICATIONS	300
APPENDIX B: REVERSE ENGINEERED SYSTEM.....	302
APPENDIX C: SAMPLE DATA	304

LIST OF TABLES

TABLE 1.1: THESIS STRUCTURE	19
TABLE 3.1: DATABASE, KEYWORDS AND ARTICLES SEARCH RESULT	48
TABLE 3.2: TOP 20 KEYWORDS IN ENERGY CONSUMPTION PREDICTION	52
TABLE 3.3: TOP 20 JOURNALS AND CITATIONS.....	53
TABLE 3.4: DESCRIPTION OF NINE POPULAR AND PROMISING STATISTICAL AND AI TOOLS	56
TABLE 3.5: DESCRIPTION OF SIX KEY CRITERIA IN THE FIELD OF ENERGY PREDICTION.....	59
TABLE 3.6: DATA PROPERTIES, PURPOSE, AND PERFORMANCE OF STATISTICAL AND AI BASED TOOLS EMPLOYED IN REVIEWED STUDIES.....	62
TABLE 3.7: SUMMARY STATISTICS OF THE ERROR TYPES REPORTED IN REVIEWED STUDIES.....	72
TABLE 3.8: MATRIX OF NUMBER OF TIMES A SET OF TWO TOOLS WERE DIRECTLY COMPARED IN THE REVIEWED STUDIES	73
TABLE 4.1: DATABASE, TERMS/KEYWORDS AND RESEARCH ARTICLES SEARCH OUTCOME.....	92
TABLE 4.2: DESCRIPTION OF NINE POPULAR AND POPULAR AND PREVALENT FEATURES OF BEP	98
TABLE 4.3: FEATURES AND ASSOCIATED RANKING	102
TABLE 4.4: DATA PROPERTIES, PURPOSE, AND FEATURES EXPLORED IN REVIEWED STUDIES... ..	103
TABLE 5.1:QUALITATIVE VS QUANTITATIVE METHOD (CASEBEER AND VERHOEF, 1997; EASTERBY-SMITH ET AL., 2001)	117
TABLE 5.2: BUILDING AND WEATHER-RELATED VARIABLES COLLECTED.....	122
TABLE 5.3: ENERGY RATING AND ENERGY PERFORMANCE VALUES	127
TABLE 6.1: RANK OF EACH FEATURE SELECTED USING VARIOUS FEATURE SELECTION METHODS	140
TABLE 6.2: RANK OF EACH FEATURE SELECTED USING VARIOUS FEATURE SELECTION METHODS	145
TABLE 7.1: PERFORMANCE RESULT FOR EACH MODEL WITH AND WITHOUT FEATURE SELECTION METHODS.	169
TABLE 7.2: PERFORMANCE RESULT FOR EACH MODEL WITH AND WITHOUT FEATURE SELECTION METHODS.	170
TABLE 7.3: FEATURE SELECTION SUITABILITY FOR CERTAIN ALGORITHMS	174
TABLE 7.4: TOP 5,6,7 & 10 FEATURES	184
TABLE 7.5: TOP 5,6,7 & 10 FEATURES	191
TABLE 7.6: MODEL PERFORMANCE ACROSS VARYING DATASET(CLASSIFICATION)	201

TABLE 7.7: MODEL PERFORMANCE ACROSS VARYING DATASET(REGRESSION)	205
TABLE 7.8: BIG-DATA MODEL PERFORMANCE.....	209
TABLE 7.9: MODEL PERFORMANCE BEFORE AND AFTER OPTIMIZATION (CLASSIFICATION)	213
TABLE 7.10:MODEL PERFORMANCE BEFORE AND AFTER OPTIMIZATION (REGRESSION)	219

LIST OF FIGURES

FIGURE 1.1: FLOW CHART DIAGRAM OF THE RESEARCH FRAMEWORK.....	16
FIGURE 1.2: FLOW DIAGRAM OF THESIS STRUCTURE	21
FIGURE 2.1: DATA FLOW AND KEY PROCESSES OF PHYSICAL MODEL SIMULATION (LI AND WEN, 2014)	27
FIGURE 2.2: FEED FORWARD NEURAL NETWORK ARCHITECTURE.	31
FIGURE 2.3: ILLUSTRATIVE DIAGRAM OF A MEDIUM ANNUAL ENERGY USE PER UNIT FLOOR	33
FIGURE 2.4: EXAMPLE OF K-NN REGRESSOR.....	34
FIGURE 2.5: DEEP NEURAL NETWORK ARCHITECTURE	35
FIGURE 3.1: FRAMEWORK OF THE PRIMARY STEPS OF THE METHODOLOGY.....	46
FIGURE 3.2: ANNUAL PUBLICATIONS OF ENERGY PREDICTION ARTICLES	50
FIGURE 3.3: KEYWORDS OCCURRENCE NETWORK.....	51
FIGURE 3.4: JOURNAL CITATION NETWORK.....	53
FIGURE 3.5: GLOBAL CO-AUTHORSHIP NETWORK	54
FIGURE 3.6: GLOBAL PUBLICATIONS AND CITATIONS DISTRIBUTION.....	55
FIGURE 3.7: THE PROPORTION OF REVIEWED STUDIES BASED ON (A) BUILDING TYPES, (B) TEMPORAL GRANULARITY, (C) ENERGY TYPES, (D) STATISTICAL OR AI-BASED TOOLS	61
FIGURE 3.8: AVERAGE RMSE AND MAE RESULT FROM STUDIES THAT CONDUCTED A DIRECT COMPARISON OF TOOLS	71
FIGURE 3.9: STACKED PLOT INDICATING THE PERCENTAGE OF STUDIES THE REPORTED ONE TOOL OUTPERFORMS THE OTHER.	74
FIGURE 3.10: AVERAGE RMSE OR MAE RESULTS FROM STUDIES THAT SPECIFIED THE BUILDING TYPE APPLIED.....	76
FIGURE 3.11: PROPORTION OF REVIEWED STUDIES THAT USED DIFFERENT BUILDING TYPES....	77
FIGURE 3.12: A SIMPLIFIED FRAMEWORK FOR TOOL SELECTION IN DIVERSE SITUATIONS	81
FIGURE 3.13: STRENGTH AND WEAKNESS OF TOOLS BASED ON REVIEW.....	82
FIGURE 4.1: FRAMEWORK OF THE KEY PHASES OF THE METHODOLOGY	91
FIGURE 4.2: PROPORTION OF ANNUAL PUBLICATION ON BEP FEATURES	94
FIGURE 4.3: KEYWORDS OCCURRENCE BIBLIOGRAPHIC MAP	95
FIGURE 4.4: TOP 20 KEYWORD AND NUMBER OF OCCURRENCES	96
FIGURE 4.5: GLOBAL COLLABORATION NETWORK	97
FIGURE 4.6: PROPORTION OF PUBLICATIONS BY COUNTRY.....	98
FIGURE 4.7: FREQUENCY OF APPLICATION OF DRIVER IN STUDY.	101

FIGURE 4.8: PERCENTAGE OF REVIEWED ARTICLES BASED ON (A) BUILDING TYPES, (B) ENERGY TYPES.....	102
FIGURE 5.1: GRAPHICAL REPRESENTATION OF THE TYPES OF BUILDINGS UTILIZED.....	121
FIGURE 5.2: MONTHLY WEATHER TEMPERATURE.	125
FIGURE 5.3: ENERGY EFFICIENCY RATING (GOV.UK).....	126
FIGURE 5.4: PROPORTION OF ENERGY EFFICIENCY RATINGS FOR EACH BUILDING TYPE.	127
FIGURE 6.1: THE FRAMEWORK OF THIS ANALYSIS.....	134
<i>FIGURE 6.2A-G: TOP 20 MOST RELEVANT.....</i>	139
FIGURE 6.3: CORRELATION BETWEEN VARIABLES (CLASSIFICATION)	141
FIGURE 6.4A-D: TOP 20 MOST RELEVANT FEATURES USING VARIOUS FEATURE SELECTION METHODS(REGRESSION).....	145
FIGURE 6.5: CORRELATION BETWEEN VARIABLES (REGRESSION).....	147
FIGURE 6.6: OLS REGRESSION ANALYSIS	149
FIGURE 7.1: FRAMEWORK UTILISED FOR MODEL DEVELOPMENT.	160
FIGURE 7.2: EXAMPLE OF ML CLASSIFICATION METHOD	164
FIGURE 7.3: FLOWCHART DIAGRAM OF THE FEATURE SELECTION IMPACT ANALYSIS	168
FIGURE 7.4: PREDICTION PERFORMANCE DISTRIBUTION FOR EACH ML ALGORITHM WITH AND WITHOUT FS METHODS(REGRESSION).....	172
FIGURE 7.5: PREDICTION PERFORMANCE DISTRIBUTION FOR EACH ML ALGORITHM WITH AND WITHOUT FS METHODS (CLASSIFICATION).....	174
FIGURE 7.6: PREDICTION PERFORMANCE DISTRIBUTION FOR ML FEATURE SELECTION METHODS	178
FIGURE 7.7: MODEL'S PERFORMANCE(R2) WITH AND WITHOUT WEATHER DATA (REGRESSION)	180
FIGURE 7.8: MODEL'S PERFORMANCE(ACCURACY) WITH AND WITHOUT WEATHER DATA	181
FIGURE 7.9: MODEL PERFORMANCE USING TOP 5 FEATURES OF VARYING FEATURE SELECTION METHOD	186
FIGURE 7.10:MODEL PERFORMANCE USING TOP 6 FEATURES OF VARYING FEATURE SELECTION METHOD.....	187
FIGURE 7.11: MODEL PERFORMANCE USING TOP 7 FEATURES OF VARYING FEATURE SELECTION METHOD	188
FIGURE 7.12:MODEL PERFORMANCE USING TOP 10 FEATURES OF VARYING FEATURE SELECTION METHOD	189

FIGURE 7.13: MODEL PERFORMANCE USING TOP 5,6,7 & 10 FEATURES OF VARYING FEATURE SELECTION METHOD	194
FIGURE 7.14: MODEL COMPARISON USING R ² AND MAE	195
FIGURE 7.15: FLOWCHART DIAGRAM OF THE PREDICTION FRAMEWORK	199
FIGURE 7.16: LINE PLOT MACHINE LEARNING MODELS' PERFORMANCE AT VARIOUS DATA SIZES	202
FIGURE 7.17: AUC SCORES OF MACHINE LEARNING MODELS' PERFORMANCE ACROSS VARIOUS DATA SIZES	203
FIGURE 7.18: RMSE SCORES OF MACHINE LEARNING MODELS' PERFORMANCE ACROSS VARIOUS DATA SIZES	206
FIGURE 7.19: R-SQUARED SCORES OF MACHINE LEARNING MODELS' PERFORMANCE ACROSS VARIOUS DATA SIZES.....	207
FIGURE 7.20: ROC CURVE FOR SVC, GB, RF, LSR	214
FIGURE 7.21: ROC CURVE FOR KNN, DT, NB, ADABOOST.....	215
FIGURE 7.22: ROC CURVE FOR ET, MLP, QDA, BAGGING	216
FIGURE 7.23: ROC CURVE FOR VOTING	218
FIGURE 7.24: MODEL PERFORMANCE FOR BOTH CLASSIFICATION AND REGRESSION	221
FIGURE 8.1: VALIDATED FRAMEWORK.....	227
FIGURE 8.2: STATISTICAL AND AI-BASED REVERSE ENGINEERING SYSTEM.....	236

ABBREVIATION

AI	Artificial Intelligence
ML	Machine Learning
HVAC	Heating, Ventilation and Air Conditioning
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
SFS	Statistical Feature Selection Method
MLFS	Machine Learning Feature Selection Method
RF	Random Forest
SVM	Support Vector Machine
DNN	Deep Neural Networks
BEP	Building Energy Prediction
GHG	Green House Gas
WHO	World Health Organisation
GB	Gradient Boosting
DT	Decision Tree
KNN	K Nearest Neighbours
XGB	Extreme Gradient Boosting Trees
UN	United Nations
EU	European Union
CO2	Carbon Dioxide
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
MLR	Multiple Linear Regression
IEEE	Institute Of Electrical and Electronics Engineers
INSPEC	Information Service for Physics Electronics and Computing
EV	Engineering Village
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
r2	R-Squared
CV	Coefficient Of Variation

RMSE	Root Mean Square Error
TBE	Total Building Energy
TE	Total Electricity
CL	Cooling
HT	Heating
NG	Natural Gas
IEA	International Energy Agency
BEPM	Building Energy Prediction Model
UK	United Kingdom
NCS	Not Clearly Stated
BIPV	Building Integrated Photovoltaic
WWR	Window-To-Wall Ratio
OLS	Ordinary Least Squares
FS	Feature Selection
MI	Mutual Information
RFE	Recursive Feature Elimination
ERF	Embedded Random Forest
ET	Extra Trees
H&C	Heating and Cooling
CY	Construction year
BFA	Building floor area
ST	Statistical Technique
SMOTE	Synthetic Minority Over-Sampling Technique

CHAPTER ONE

1.0 INTRODUCTION

1.1 Chapter Overview

This chapter provides a comprehensive overview of this research by discussing the background, aim, objectives, hypotheses, and contributions. The background section describes building energy consumption and its environmental implications. Following this, the aim and objectives are highlighted to convey the goals and directions of the research. The hypotheses section presents the expected experiment and analysis conducted throughout the research. While the research methodology section outlines the approach and methods employed to investigate the research questions, the research framework also offers a conceptual structure guiding the investigation. Subsequently, this chapter presents the contributions this research aims to engender in the field of building energy consumption. Justification for the research is also provided, expounding the significance and relevance of this research. Moreover, the scope of the research is outlined, displaying the coverage of the research. Finally, the thesis structure is summarised, highlighting chapters disseminated through academic publications.

1.2 Background

Energy-efficient and sustainable buildings have become imperative towards saving the environment, as building inefficiency significantly contributes to global energy consumption and greenhouse gas emissions (Pham et al., 2020). The high proportion of energy consumed in buildings leads to major environmental problems causing climate change, air pollution, and thermal pollution, among others, which deploys a severe impact on the existence of mankind (Dandotiya, 2020). In the past decades, energy demand in buildings has amplified considerably due to population increase and prompt urbanization (Aversa et al., 2016). This rise in energy consumption and its related adverse effects prompts the importance of understanding building energy efficiency and exploring various potentials for enhancement, so new buildings can be developed to consume energy more efficiently (TOPRAK et al., 2017). Research indicates that ensuring future buildings are more energy efficient requires a focus on energy efficiency at the design stage (Ding et al., 2018; Jaber and Ajib, 2011; Wang et al., 2014). Design stage decisions

can considerably impact the building energy performance, with potential energy savings of around 30% achievable during this stage (Aversa et al., 2016; Colmenar-Santos et al., 2013).

Customarily, energy assessment at the design stage is achieved through energy simulation models. Building designers currently rely on building energy simulation tools which often require a large number of parameters such as HVAC (Heating, Ventilation and Air Conditioning) system, insulation thickness, building thermal properties, internal occupancy loads, and solar information, among others (Runge and Zmeureanu, 2019), and consumes a lot of time (ranging from a few hours to several days) to process the potential energy performance of buildings (Jhai, 2023; Pham et al., 2020). These simulation tools include mainly DOE-2, EnergyPlus, and TRNSYS, among many others. Considering the amount of time taken to deduce the relatively high number of parameters and utilize simulation tools, it is considered inefficient (Pham et al., 2020). Occasionally, building designers optimize designs based on previous experience which often results in inaccurate conclusions (Deb et al., 2017a; Runge and Zmeureanu, 2019; Wang et al., 2005). It is stated that energy prediction models have promising potential for significant energy conservation and better decision-making in the optimization of building design to decrease total energy usage (Hamed and Nada, 2019). Given the potential benefits of building energy consumption prediction and the need to better understand building energy efficiency, energy consumption prediction has captivated the attention of many researchers, which has generated new developments with the utilization of diverse techniques for operational buildings (Ahmad et al., 2017a; Fathi et al., 2020a; Serale et al., 2020; Bourdeau et al., 2019; Runge and Zmeureanu, 2019; Amasyali and El-Gohary, 2018).

Building energy consumption prediction is performed through the utilization of data-driven models(Ahmed Gassar et al., 2019a; Akbar et al., 2020), which require less number of building parameters. Data-driven models are based solely on mathematical models and measurements. This model utilizes statistical and Artificial Intelligence (AI) based algorithms for building energy prediction. AI/Machine Learning (ML) method is recognised as contemporary and one of the most effective techniques in the energy prediction domain (Ríos Canales, 2016; Vorobeychik and Wallrabenstein, 2013) In comparison to simulation tools also known as physical or forward models, AI algorithms simply consume building energy-related data to predict energy consumption and it does not require a sizeable number of detailed inputs about the building. Also, data-driven methods have been proven to be more accurate and efficient, as the result can be generated in seconds (Runge & Zmeureanu, 2019; Qiao, Yunusa-Kaltungo

and Edwards, 2020). For example, Neto & Fiorelli (2008) conducted a comparative analysis of a data-driven method and physical simulation method for predicting building energy use. It was concluded that the data-driven method outperforms the physical method. In recent years, artificial intelligence (AI) algorithms have been extensively employed and produced good outcomes in the field of building energy consumption prediction (Aversa et al., 2016, 2016; K. Li et al., 2018; Pham et al., 2020; Qiao et al., 2020a; Caleb Robinson et al., 2017a). Researchers propose that the availability of a building energy prediction model with accurate predictions is expected to save around 30% of total energy consumption in buildings (Aversa et al., 2016; Colmenar-Santos et al., 2013). Hence, the continuous effort to improve building energy use prediction accuracy is essential for more energy-efficient buildings. However, not many studies have explored its suitability for prediction at the early design stage of buildings.

There are several applications of ML algorithms in the field of energy consumption with the goal of developing an accurate model for operational buildings, yet there is less focus on forecasting energy consumed at the early design stage. Despite the need for accurate predictive models to better understand building energy efficiency, several aspects still require further investigation towards the development of an accurate building energy consumption model. Understanding the factors influencing energy consumption and prediction performance is essential to potentially improve the accuracy of a model. At the design stage, the development of a prediction model with excellent performance would provide great support for building designers. It will enable designers to deduce the potential energy use of a building at the design stage, optimize design instantaneously based on the energy predicted and conduct continuous iteration until optimum performance is achieved. This will potentially reduce the construction of energy-inefficient buildings. In response to this need, this research developed an AI/ML model for predicting potential energy consumption at the design stage and goes a step further to streamline the iteration process and enhance efficiency. This research developed a reverse engineered system that enables building designers to specify the target energy consumption value and receive the optimal values for each feature, thus limiting the continuous iteration of features to achieve target energy consumption. This would potentially decrease the construction of more energy-inefficient buildings that are detrimental to the environment.

1.3 Research Problem and Gap in Knowledge

AI/ML algorithms are recognized as one of the most effective tools for prediction tasks (Ríos Canales, 2016; Vorobeychik and Wallrabenstein, 2013). However, the accuracy of these

models in predicting building energy consumption depends on factors such as model selection, and the quantity and quality of the data (Runge and Zmeureanu, 2019). Numerous studies have compared various AI/ML algorithms using different quantities of data (Dong et al., 2021a; Cheng Fan et al., 2017; Pham et al., 2020; Wang et al., 2020). For example, in 2019, Runge and Zmeureanu investigated the use of artificial neural networks (ANN) for predicting hourly building energy consumption. They found that the ANN, when trained on a single commercial building dataset, yielded poor hourly predictions, suggesting that the performance might be limited by the dataset used (Runge and Zmeureanu, 2019). Bagnasco et al. (2015) applied ANN to predict electrical consumption in a hospital building using weather-related data and time/day variations, achieving better results in winter (Bagnasco et al., 2015). In the ML field, it is generally hypothesized that larger datasets improve model performance and result reliability (Dalal, 2018; Goyal et al., 2020; Kabir, 2020; Kaur and Gupta, 2017; Lee et al., 2011). While researchers in other fields have previously noted the challenges of using large datasets to develop models and employ big data analytics (Alaka et al., 2019; Balogun et al., 2021; Swanson and Xiong, 2018), the advancement has not been extensively applied in energy consumption prediction. Despite the good performance of ML algorithms shown in past studies (e.g. Aversa et al., 2016, 2016; K. Li et al., 2018; Pham et al., 2020; Robinson et al., 2017), it is suggested that using even larger datasets would further enhance performance (Runge and Zmeureanu, 2019).

These algorithms have been very prominent in research, due to their relatively good performance in energy prediction. However, the selection of these tools for exploration or investigation is often done arbitrarily or based on popularity (Divina et al., 2018a; Feng and Zhang, 2020; D.M.F. Izidio et al., 2021; C. Robinson et al., 2017); except for a few articles(e.g. Culaba et al., 2020). The inefficient method of tool selection often leads to poor model performance and time-consuming comparative analyses of tools, rather than utilizing and optimizing the appropriate tool for the specific situation. In research, this tool selection method can be due to a lack of adequate evaluation reports of data-driven energy prediction tool performance, centred on diverse pertinent conditions or situations. It is evident that the performance of several models, such as building energy prediction models, is greatly contingent on the tool selected, and features selected among other factors (Goyal et al., 2020; Kabir, 2020; Olu-Ajayi et al., 2022a; Runge and Zmeureanu, 2019).

When evaluating the best algorithm, it is clear that the accuracy results of algorithms applied to different datasets are not directly comparable, as each dataset and situation produces different results (Demsar, 2006). This research conducts a rigorous comparative analysis of several algorithms identified by means of a systematic literature review, applying them to the same data and situation to ensure a fair comparison. Despite the significant importance and widespread use of various tools for predicting building energy consumption, there is no consensus on the best or most suitable tool, and current research primarily focuses on operational buildings. Developing a highly effective energy prediction model for the design stage is crucial for preventing the construction of energy-inefficient buildings. However, this development depends on several factors that have not been thoroughly addressed in the literature. The hypotheses section (1.3.2) below further elucidates these problems and gaps aligned with the specific research objectives.

1.4 Research Aim and Objectives

The primary aim of this research is to develop a reverse-engineered system for efficient energy assessment at the design stage of residential buildings. This allows for a back-to-front method where a building designer can simply specify the desired or target energy value for a design, input it into the model, and receive optimal value for each building feature, essential to achieve the desired or target energy value. To accomplish the stated aim, the following objectives were structured as follows.

1. To establish the key features that influence energy consumption in buildings using a systematic literature review.
2. To establish and validate the most prominent statistical and AI/ML algorithms for building energy consumption prediction in literature, using a systematic literature review.
3. To establish the minimum data size required for developing an efficient AI/ML energy prediction model, by benchmarking the performance of ML algorithms with varying data sizes.
4. To identify the most accurate and efficient building energy prediction model and the best-performing transparent statistical and AI/ML models

5. To develop an optimization model using the best performing transparent model from objective 4.

1.4.1 Research Questions

To fulfil the aim and the objectives outlined above, the following questions have been formulated to guide the execution of this research:

1. What are the most common features that influence energy consumption in buildings?
2. What are the most prominent statistical or AI/ML algorithms for energy consumption prediction?
3. What is the minimum data size required for efficient energy prediction at the design stage of buildings?
4. Which statistical or AI/ML algorithm is the most accurate, efficient, and transparent for predicting potential energy consumption at the design stage of buildings?
5. What are the key components and parameters necessary to develop an optimization model?

1.4.2 Research Hypothesis

The following hypotheses are proposed in relation to the aforementioned research objectives in section 1.3. With respect to objective #1, the importance of understanding features influencing energy consumption at the design stage cannot be overemphasized, as the design stage is where the potential lies for around 30% energy savings (Aversa et al., 2016; Colmenar-Santos et al., 2013). The identification of the most relevant features at the design stage will aid the development of high-performing building energy prediction models that will be advantageous for designers to create energy-optimal designs. Thus, chapter four of this research delivers a holistic, structured, and comprehensive review of studies that have explored various features affecting energy use in buildings, to establish the factors that have been commonly selected as important based on analysis. Furthermore, beyond the theoretical identification of features that guide feature selection for the development of energy prediction models, it is essential to conduct practical experiments to identify relevant features based on the dataset utilized. From the literature, some of the most important features identified include openings (such as windows and doors), some of the claims of the literature vary according to the dimension of the effect on building energy consumption.

key features

For example, the weakest component of the building envelope, accountable for the most significant amount of heating and cooling energy consumption has been established as windows (M. Alwetaishi and Benjeddou, 2021; Chiesa et al., 2019; Yoshino et al., 2017). However, some other research (e.g., Ihara et al., 2015; Park et al., 2020; Rouleau et al., 2018) proclaims that outdoor temperature has the greatest effect on energy consumption in buildings. Hence, it is proffered that proper insulation of walls, and roofs, among others have the most significant effect on the heating or cooling energy consumption of the building depending on the weather conditions (L. Y. Zhang et al., 2017). Additionally, Florides et al., (2002) stipulated that roofs are the most essential structural components of the building that engender 19% of energy savings when properly insulated. The assumption that openings in a building have the most significant effect on energy consumption is relatively popular in the field of building energy evaluation (Chiesa et al., 2019; Yoshino et al., 2017), however, there are equally a good number of studies that have established other areas as highlighted above. While openings like windows and doors can contribute to heat gain or loss, some studies argue that it does not have the most significant effect on energy consumption in buildings (Ihara et al., 2015; Park et al., 2020; Rouleau et al., 2018). In this regard, the following hypothesis has been proposed based on literature:

H0: Openings (such as windows and doors) have the most significant effect on building energy consumption.

One experimental approach that can be employed to achieve H0 is feature selection, which involves identifying the most relevant features that impact building energy consumption. This can provide more valuable insights, as this will allow comparison between different building features, and weather features, which can further solidify the understanding of the impact of building openings on energy consumption in different contexts. Previous studies have employed this approach in both classification (Bahassine et al., 2020; Bommert et al., 2020; Iqbal et al., 2020) and regression prediction (Dong et al., 2021b; Fan et al., 2014a). While some studies have argued that feature selection is more effective in classification than regression prediction (Jović et al., 2015; Kumar, 2014), several studies have employed feature selection for regression prediction and achieved outstanding performance (Ahmad et al., 2017b; Fan et al., 2014a; Kolter and Ferreira, 2011; Paudel et al., 2017). For example, Paudel et al., (2017) employed feature selection for regression prediction of building energy consumption and achieved an accuracy of 98% (R-squared). It was emphasized that the benefits of feature selection for regression prediction go beyond increasing accuracy levels, it also improves

computational efficiency. However, Kapetanakis et al., (2017) doubted this and conducted an extensive experiment to ascertain the veracity of the theory. In his work, he performed a comparative analysis of regression prediction models for predicting the thermal load with and without feature selection. It was established that the accuracy of regression prediction remained static with and without the application of feature selection. Notwithstanding, Faisal et al., (2019) conducted a similar experiment for regression prediction of electricity consumption in buildings and it was concluded that the results produced using the selected features outperformed the results without the application of feature selection. Therefore, there is no clear agreement on the effect of feature selection on the accuracy of a regression model. Thus, the following hypothesis is proposed:

H1: Feature selection does not have an effect on regression model performance.

Hsu, (2015) stipulated that one of the most broadly unaddressed issues in energy consumption literature, which affects machine learning algorithms performance is feature or variable selection. Some studies have applied the feature selection method in the development of energy predictive models (M. W. Ahmad et al., 2017; Z. Dong et al., 2021; Zhang & Wen, 2019b). However, the fraction of studies that deliver comprehensive insights into the incorporation and capabilities of feature selection with machine learning is still limited (Olu-Ajayi et al., 2023b). Generally, beyond the identification of the most relevant features, feature selection proffers great advantages when properly applied to machine learning prediction (Zhang and Wen, 2019). However, notwithstanding the noted importance of feature selection in machine learning prediction according to a majority of studies (e.g., Ahmad et al., 2017; Dong et al., 2021a; Zhang and Wen, 2019 among many others), it has been argued by a few studies that carried out experimental research that feature selection can have negative impacts on ML model performance (Balogun et al., 2021; Kapetanakis et al., 2017). For example, Balogun et al., (2021) conducted an experiment to investigate the effect of Boruta feature selection on machine learning models for air pollution prediction. While it was noted that random forest and decision tree produced better performance without feature selection, all other algorithms produced better performance based on r-squared such as Artificial neural network which produced 77 percent with feature selection and 71 percent without feature selection. Therefore, it is hypothesised that:

H2: Feature selection has positive impacts on machine learning energy prediction model performance.

A critical review of several studies that employed different feature selection methods (e.g., Ahmad et al., 2017; Dong et al., 2021a; Zhang and Wen, 2019 among many others) suggests that while feature selection does have an impact on prediction performance, this impact could be negative or positive depending on the feature selection method utilized. Feature selection methods can generally be classified into two broad categories: machine learning feature selection method (MLFS) and statistical feature selection method (SFS). MLFS methods are ML algorithms employed for ranking relevant features in the development of a prediction model (e.g., random forest (Huljanah et al., 2019; Nguyen et al., 2013), Extratrees (Sharma et al., 2019), among others). SFS methods are statical methods utilised to identify the most relevant features for the development of a prediction model (e.g., Chi-square (Bahassine et al., 2020; Jin et al., 2006), Analysis of variance(ANOVA) (Ding et al., 2014), among others)

It was noted that the majority of studies that highlighted the positive impact of feature selection employed the machine learning feature selection method(e.g., Ahmad et al., 2017; Dong et al., 2021a, 2021a; Zhang and Wen, 2019), For example, Guo et al., (2020) applied Boruta feature selection in the development of energy prediction models using random forest(RF), and Support Vector Machine(SVM) and produced good accuracy of 82% and 90% respectively. Similarly, Nguyen et al., (2013) achieve 99% accuracy by applying RF for feature selection and model development. While the majority of the few that argued against, utilized the statistical feature selection method(e.g.,Bahassine et al., 2020; Chou and Bui, 2014; Kapetanakis et al., 2017) For example, Chou and Bui, (2014) applied chi-square feature selection for predicting building energy consumption using RF and achieved 65.9% accuracy. However, Szul et al., (2021) applied the machine learning feature selection method (Boruta) for energy use prediction and noted it had a minute effect on the model performance. Nonetheless, a very recent study, Kanyongo and Ezugwu (2023) also emphasised that machine learning-based feature selection is more effective than the statistical feature selection methods for machine learning prediction. Therefore, it is hypothesised that:

H3: ML feature selection methods lead to better performance of machine learning prediction models than statistical feature selection methods.

Machine learning feature selection methods are embedded methods premeditated to pinpoint the most relevant features for the development of ML prediction models which is expected to engender more accurate and efficient models (Newman et al., 2022). For example, the random forest feature selection method was premeditated for the identification of the most relevant

features that engender the best performance when random forest or other ML algorithms are used for machine learning prediction. Statistical feature selection methods, on the other hand, are not designed for a specific algorithm, and may not effectively identify the most relevant features (Pudjihartono et al., 2022). However, they are often able to identify features that are generally relevant and the ones that are generally very irrelevant without recognizing features that can be relevant when combined with other features (Bommert et al., 2020), thus still leading to improved model performance to some extent across various algorithms.

Although MLFS methods have been proffered to be most suitable for better ML model performance in many studies (M. W. Ahmad et al., 2017; Z. Dong et al., 2021; Zhang & Wen, 2019b), some studies have argued that they do not perform best when using the same algorithm for feature selection and model development. For example, Ahmad et al., 2017a utilized random forest feature selection in the development of Random Forest (RF) and Artificial Neural Networks (ANN) building energy prediction models. The prediction model with the application of random forest feature selection produced better performance in ANN than RF. Similarly, Guo et al., (2020a) applied the Boruta (built around random forest algorithm) feature selection method in the development of RF and SVM building energy prediction models. SVM produced 90% accuracy while RF produced 82% accuracy. However, some experimental studies contend that the application of random forest for feature selection and model development can significantly improve model performance (Huljanah et al., 2019; Nguyen et al., 2013). For example, Nguyen et al., (2013) random forest for feature selection and model development in predicting prostate cancer achieved 99% accuracy. Therefore, it is hypothesized that:

H4: Using the same algorithm for feature selection and prediction leads to better model performance than using different algorithms.

As stated in H0, several studies have stipulated that outdoor temperature has the greatest effect on energy consumption in buildings (Ihara et al., 2015; Park et al., 2020; Rouleau et al., 2018). Research by Rouleau et al., (2018) established that weather features such as humidity, precipitation, and outdoor temperature among others have an effect on heating and cooling loads. Various studies have employed weather data for developing machine learning models for predicting energy consumption and noted good model performance(e.g., Bagnasco et al., 2015; Ding and Liu, 2020; Dong et al., 2005, 2021b, p. 2; Kim et al., 2020 among others). For

example, Kamel et al., (2020) utilized weather data and eXtreme Gradient Boosting(XGB) algorithm for predicting building energy consumption and achieved a Root mean square error(RMSE) of 0.038. Lee and Rhee, (2021) also employed for Artificial Neural Network (ANN) for predicting building energy consumption using weather data and achieved 0.27 RMSE. Generally, less than 0.3 RMSE values are considered good performance for a prediction model (Asilevi et al., 2019; Ihsan, 2020; Rupf et al., 2021; Shen et al., 2022). However, other studies that did not employ weather data for developing machine learning models for predicting energy consumption have achieved similar performances. For example, Almalaq and Zhang, (2019) achieved 0.186 RMSE using ANN for predicting energy consumption without weather data. Similarly, Izidio et al., (2021) achieved 0.077 RMSE using SVM for building energy consumption prediction. Hence, it cannot be claimed that weather data has a significant effect on ML model performance. Also, none of the studies conducted a clear and unbiased comparison of the model performance with and without weather data to validate the effect of weather data on model performance. Nonetheless, studies such as Karatasou et al., (2006) utilized feed-forward neural networks for predicting hourly energy consumption in buildings using two different datasets consisting of various weather variables(i.e., temperature, solar radiation wind humidity). Each dataset was used to develop a neural network model for further performance analysis. It was found that weather data variables (i.e., wind velocity or humidity) have a less significant effect on neural network model performance. Clearly, weather data has an effect on energy consumption, thus it can be argued that certain ML algorithms do not effectively capture the effect of weather data in predicting energy consumption in buildings. Therefore, it is hypothesized that:

H5: Weather data do not significantly improve ML model performance for building energy consumption prediction.

With respect to objective #2, It is well noted that the performance of diverse statistical and AI/ML tools is highly predicated on the method and algorithm selected respectively, among other factors (Goyal et al., 2020; Kabir, 2020; Runge and Zmeureanu, 2019). In numerous AI/ML building energy prediction (BEP) studies, algorithm selection is centred more on popularity (e.g., Divina et al., 2018; Fan et al., 2017; Feng and Zhang, 2020; Somu et al., 2021) than its capabilities. For example, without much justification, Ding et al., (2021) utilised ANN for building energy consumption prediction models using data from only 1 building and a sample size of 8760 and ANN produced poor performance of 31.98 RMSE. Meanwhile, it is noted that SVM thrives better in small datasets (Mat Daut et al., 2017), while ANN is more

dominant in large datasets which enables the neural network sufficient data to train the model (Bourhnane et al., 2020). (see more in chapter 3.8.6). It is important to conduct empirical experiments on these tools for predicting building energy consumption. Many studies have conducted a comparative analysis of the performance of different Statistical, or AI/ML-based tools for energy consumption prediction(Ahmed Gassar et al., 2019a; Alduailij et al., 2021; S. Cho et al., 2019; Liao et al., 2020; Lin et al., 2021; Parhizkar et al., 2021). However, AI and Statistical based tools have at different times outperformed each other. For example, Somu et al., (2020) compared SVM(ML) and ARIMA(Statistical) for predicting energy consumption and it was noted that SVM produced better performance. Also, in the comparative study by Guo et al., (2018), LR(Statistical) outperforms SVM for predicting energy consumption. A careful review of these studies shows that ML produce better performance in the majority of the studies(Kontokosta and Tull, 2017; Li and Yao, 2020; Sha et al., 2019). Therefore, it is hypothesized that:

H6: Machine learning algorithms produce better performing building energy prediction performance models than statistical methods.

In recent years, deep learning algorithms have shown great promise in various fields (Abrol et al., 2021; Brinker et al., 2019; Hekler et al., 2019). It has become prominent based on the production of good performance in various other studies (Almalaq and Zhang, 2019; C. Fan et al., 2017; Somu et al., 2021). For example, Sadeghi et al., (2020) conducted a comparative analysis of different machine learning algorithms and Deep neural networks (DNN) produced the best performance with a mean absolute error value of 1.22. However, some studies such as (Amber et al., 2018; Köhler et al., 2021) stipulated that deep learning does not outperform classical machine learning. For example, Ding et al., (2021) conducted a comparison between artificial neural networks(ANN) and Deep neural networks(DNN). It was noted that ANN produced better performance than DNN with mean absolute error values of 11.86 and 24.46 respectively. Likewise, the experiment conducted by Liao et al., (2020), shows that random forest(RF) outperformed DNN with an r² of 88% and 70% respectively. Nonetheless, Amber et al., (2018) highlighted that the performance of DNN is dependent on certain factors. For instance, they are not as favourable in studies with a limited amount of data thus their performance relies heavily on a large amount of data. The application and evaluation of deep learning and classical machine learning for building energy prediction is vital as it holds the potential for more accurate and reliable models. Therefore, it is hypothesized that:

H7: Deep learning algorithm outperforms classical machine learning algorithms.

With respect to objective #3, it is stipulated that the accuracy of statistical/ML algorithms for predicting building energy consumption is dependent on these three factors: algorithm selected, quantity and quality of the data (Runge and Zmeureanu, 2019). It is important to conduct a comparative analysis of the effect data size has on the prediction model performance and the minimum data size required to achieve a satisfactory model performance. It is a popular hypothesis in the machine learning world “The larger the data, the more accurate the result” (Dalal, 2018; Goyal et al., 2020; Kabir, 2020; Kaur and Gupta, 2017; Lee et al., 2011). Some studies have tested this hypothesis and concluded that large data size does engender improved model performance. For example, Ngarambe et al., (2020) stipulated that Deep Neural Networks (DNN) produce good model performance on large datasets but produce poor model performance on smaller datasets. Amber et al., (2018) also noted that DNN models are not as favourable in studies with a limited amount of data thus their performance relies heavily on a large amount of data. However, some studies such as (Mat Daut et al., 2017) have disputed that large data sizes do not generally lead to better performance and some ML algorithms such as Support Vector Machine (SVM) thrive better in small datasets. More recently, studies have demonstrated that SVM has the capacity to produce good outcomes regardless of the data size (Aversa et al., 2016; Mat Daut et al., 2017; Olu-Ajayi et al., 2021). Although majority of the studies align with the popular hypothesis(Dalal, 2018; Goyal et al., 2020; Kabir, 2020), it can be argued that this hypothesis is only true for specific ML algorithms such as DNN. Considering ML algorithms rely on historical data to learn patterns and make predictions (Somu et al., 2021). It is proffered that a larger dataset provides the ML algorithm with more information and samples to learn and elicit better performance(Dalal, 2018). If the effect of large data sizes actually varies depending on the algorithm utilized, it is imperative to evaluate the impact of data size on various algorithms. This will limit the time-consuming comparative analysis of tools in several studies and guide decision-making in situations where data availability is limited(Alawadi et al., 2020; Demsar, 2006; Ding and Liu, 2020; Olu-Ajayi et al., 2023a). Therefore, it is hypothesized that:

H8: Larger data size only improves the model performance for certain ML algorithms.

With respect to objective #4, the facilitation of energy assessment at the design stage will curtail the construction of energy-inefficient buildings, as building designers optimize designs accordingly and effectively. Current traditional methods of energy assessment at the design stage are deemed inefficient due to the time consumption and labour intensity (Deb et al., 2017;

Runge & Zmeureanu, 2019). In contrast, statistical and ML algorithms are a potentially more efficient alternative(Adegoke et al., 2022; Bustos et al., 2022). In various fields such as healthcare (Jin et al., 2006; Leyh-Bannurah et al., 2018; Zheng et al., 2019), pollution prediction (Balogun et al., 2021; Sulaimon et al., 2021), and bankruptcy prediction(Alaka et al., 2018; Barboza et al., 2017; N. Wang, 2017), among others, statistical and machine learning algorithms are producing good performance in terms of accuracy and computational efficiency. However, some studies (Feurer and Hutter, 2019; Kumar et al., 2020; Smith et al., 2019)) have noted that these models can be computationally inexpensive. For example, Particularly in the field of energy consumption prediction, Ilager et al.(2021) stated that ML models can be inaccurate and computationally expensive. Nevertheless, Amasyali and El-Gohary, (2018) stated that some studies such as decision trees and statistical algorithms are generally computationally inexpensive. Therefore, it is hypothesised that:

H9: Statistical/ML tools can assess energy consumption faster than the traditional method.

If this hypothesis holds true, the incorporation of a Statistical/ML model for energy consumption prediction at the early design stage will engender timely decision-making and prompt identification of energy-saving opportunities.

1.4.3 Research Methodology

In this research, the quantitative method was selected using the positivist approach, this helped accomplish the aim as ML algorithms utilize quantitative data (numerical values) for prediction. Also, it is considered the most suitable because it focuses on specific behaviours that can be quantified and does not manipulate variables (Cozby and Bates, 2012). Qualitative research was not chosen because it is the analysis of qualitative data (text data). Qualitative and quantitative approaches vary in different ways (i.e., how data is collected, the nature of the data, the method of analysing the data and interpreting the results) (Haas, 2002). The nature of qualitative research is constructed as looking through a wide lens to discover patterns of correlation between an earlier unspecified set of concepts, while quantitative research looks through a narrow lens at a specified set of variables (Brannen and Coram, 1992). To complete the objectives, academic literature will be reviewed to identify the common and relevant features that influence building energy consumption in the United Kingdom (UK). A systematic review of existing academic literature will be conducted to identify prominent statistical and AI/ML algorithms for predicting building energy consumption. This research

will be based on the development of a statistical and AI/ML prediction for energy assessment at the design stage of buildings.

1.5 Research Framework

The framework for this research is visualized in Figure 1.1 below.

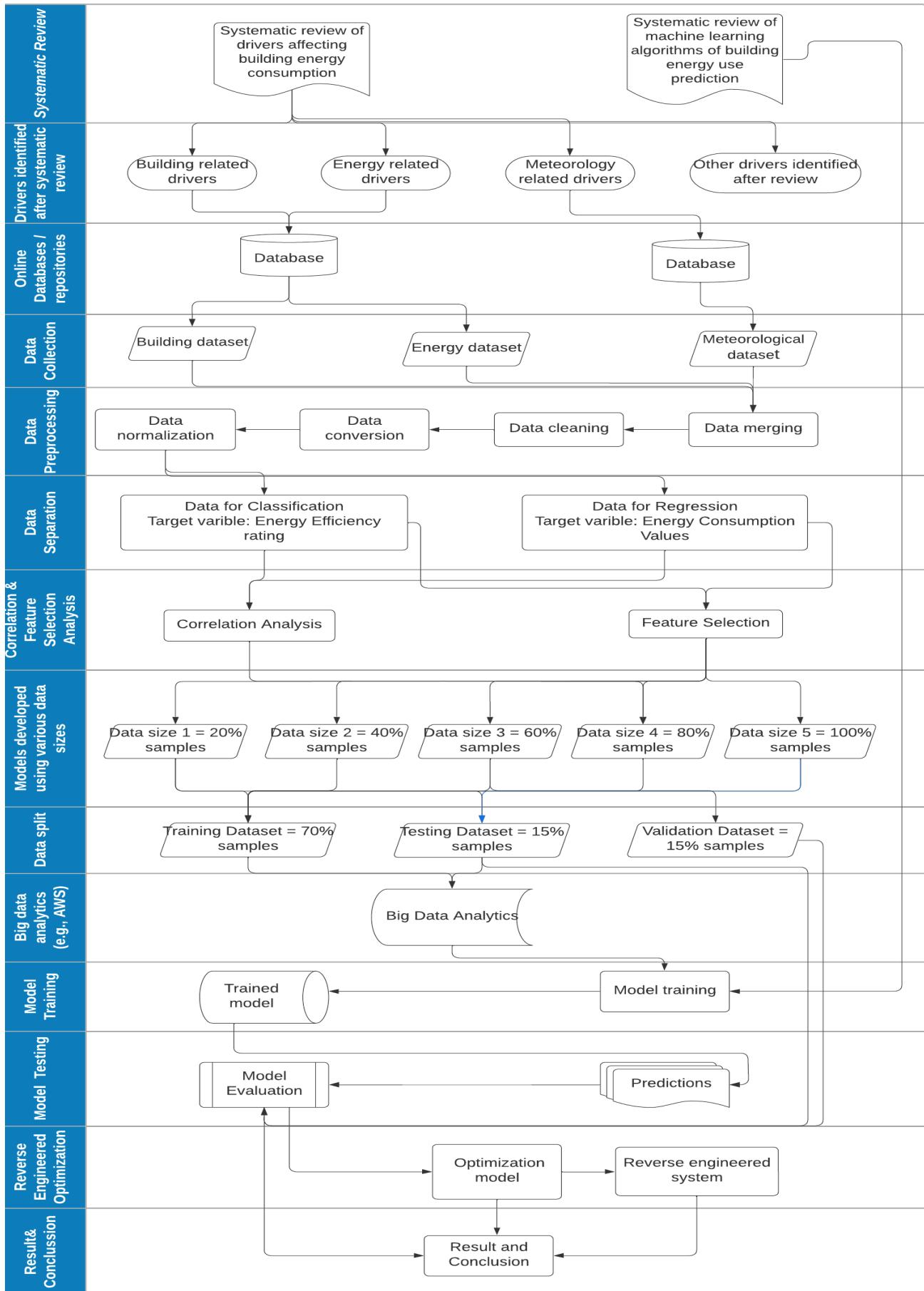


Figure 1.1: Flow chart diagram of the research framework.

1.6 Research Contributions

By completing this research study, contribution to academia and the industry sector are anticipated. This research will provide meaningful academic contributions to the prediction of building energy consumption. This research proffers several contributions by addressing the aforementioned hypotheses and objectives. This will be further elucidated in the respective sections of this research. However, this research developed a reverse engineered system that enables building designers to receive optimal value specifications for each building feature to achieve the stipulated target energy value. Traditional tools, are both resource-intensive and time-consuming, often taking over an hour to complete a single simulation. In contrast, this research focuses on leveraging advanced statistical and AI tools to streamline the energy assessment process during the design stage. The research conducted a comprehensive comparative analysis of statistical methods and AI techniques to develop an effective energy assessment model. By integrating these advanced tools, the research seeks to overcome the limitations of previous research, which has not fully explored the potential of AI in this context. AI-based models are noted for their better accuracy and speed, significantly enhancing the efficiency of the design process. One of the key advantages of using an AI model at the design stage is the ability to perform faster iterative processes. However, this research goes a step further by reducing the iteration time even more. The innovative aspect of this research is the introduction of a reverse-engineered data-driven system that allows designers to input values for each feature and the desired or target energy consumption value into the system. Once the AI model predicts the potential energy consumption based on the design feature values supplied and returns the optimal values for each design feature. This approach minimizes the need for repeated iterations, as the system directly provides the adjustments needed to meet the target energy consumption. This not only saves time but also enhances the precision and effectiveness of the design process. Concisely, this research seeks to revolutionize the energy assessment process at the design stage of a building. Additionally, contrary to previous works that focus on comparing a few ML algorithms on a relatively small dataset, this research will perform a comparative analysis of prominent and efficient ML algorithms on a large dataset by leveraging big data analytics.

1.7 Justification of Studies

By completing this research study, the contribution envisaged for the industry and academia with the development of an energy predictive model for the early design stage of building, leading to a decrease in the construction of more energy-inefficient buildings. Hence, this model will essentially improve decision-making and regulate energy consumption. Furthermore, the utilization of the reverse engineered system will lead to a reduction in energy consumption and subsequently, aid the UK in achieving its long-term target to decrease the rate of CO₂ emission by 26% and 80% in the years 2020 and 2050 respectively (Department of Energy and Climate Change, 2009). Building designers currently rely on previous experience or physical models (building energy simulation tools), which usually require large parameters and consume a lot of time (ranging from a few hours to several days) to determine the energy performance of a building. This system will certainly help expedite the work of designers by simply inputting only ten design features required by the AI/ML model and obtain a result in seconds to optimise the building design appropriately and utilize back-to-front method by simply inputting desired or target energy consumption value required by the reverse-engineered system and obtain a precise value for each feature of the building design.

1.8 Research Scope and Limitations

This research focuses on acquiring a large number of energy data, building metadata and meteorological data for buildings within the UK which would require the use of big data analytics (e.g., AWS or Microsoft Azure). The maximum data size used in this research is larger than the data used in the literature. Generally, this research will be limited to residential buildings and annual energy consumption. This research will be limited to statistical and machine learning (ML) algorithms and features established in academic literature. This research will collect data mainly only from buildings within the UK.

1.9 Thesis Structure

This section presents an overview of this research and a brief description of each chapter in Table 1.1 and Figure 1.2 below. This thesis consists of 4 research articles in appendix A and eight chapters as illustrated below:

Table 1.1: Thesis Structure

Chapter Number	Chapter Title	Summary
Chapter 1	Introduction	This chapter provides a comprehensive overview of this research by discussing the background, aim, objectives, hypotheses, and contributions.
Chapter 2	Overview of Building Energy Consumption Prediction: Methods, Simulations, Techniques, Theory and Performance Factors	This chapter will provide a review of existing literature and highlights the research gaps. It also reviews the theory of current practice.
Chapter 3	Systematic literature review and theoretical framework: Statistical and artificial intelligence-based tools for building energy consumption prediction	This chapter will present a detailed systematic review of prominent and efficient Machine Learning (ML) algorithms in the field of building energy prediction.
Chapter 4	Systematic literature review: Analysis of features influencing building energy performance	This chapter will provide a systematic review of academic literature of the most common and relevant features that influence energy consumption of buildings.
Chapter 5	Research Methodology	This chapter will present the procedures, research paradigms, design and research strategies that are adopted to achieve this research study. It will also discuss the data collection and analysis employed in this research.
Chapter 6	Feature selection evaluation for building energy consumption prediction	This chapter investigates the effectiveness of feature selection centred on building energy consumption prediction. This chapter examines the most relevant features by analysing different feature methods. The identification of the best feature combination has the potential to enhance model performance, reduces

		computational complexity, and mitigates the risk of overfitting.
Chapter 7	Development of AI/ML building energy consumption prediction model	This chapter investigates the impact of data size on performance of building energy consumption prediction models. The identification of the optimum data size required to achieve satisfactory model performance. This chapter examine the effect data size in both classification and regression prediction model performance and delivers the theoretical and practical implications of such analysis.
Chapter 8	Reverse engineered system and validated framework for building energy prediction	This chapter delivers a comprehensive validated framework for energy consumption prediction based on the investigation and outcomes of the experiment. The insights extrapolated from empirical experiments led to the creation of a validated framework. This chapter also delivers the development of the reverse-engineered system
Chapter 9	Conclusions and recommendations	The chapter concludes this research and provides a comprehensive overview of the research study. This chapter will review the research objectives, summarize key aspects such as main findings, contributions, limitations, and recommendations for future research. By reviewing these elements, this chapter aims to capture the substance of the research journey and its implications on study and practice.

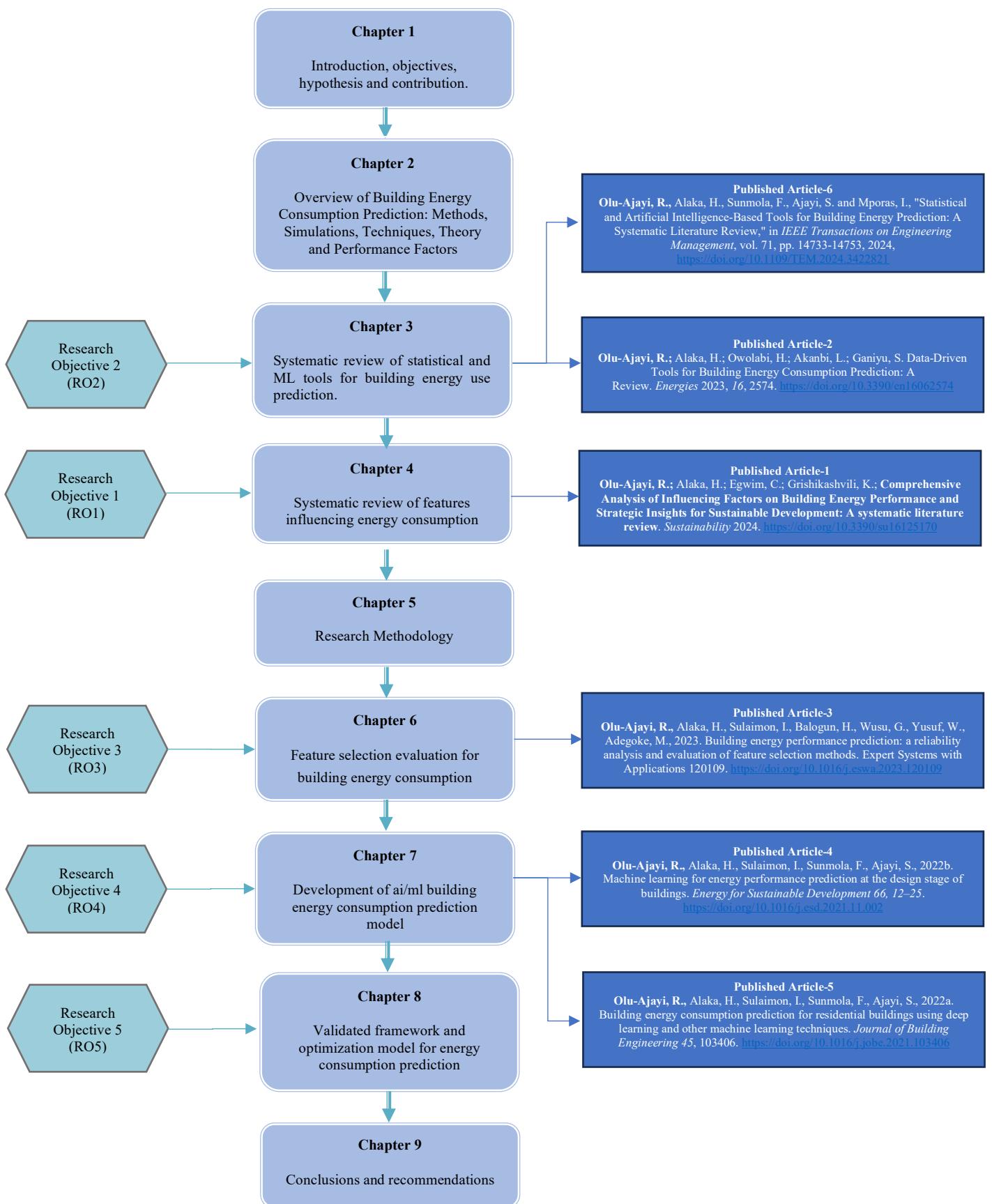


Figure 1.2: Flow Diagram of Thesis Structure

1.10 Chapter Summary

In summary, this chapter introduced the research with a clear delineation of the research problem, the background of the research, research aim and contributions. This chapter discussed energy consumption and its detrimental impact on the environment. This chapter introduces the various research hypotheses that will aid fulfilment of the research aim and objectives. The research scope is restricted to data from buildings in the UK. The fundamental research contribution is the development of a reverse-engineered system that enables building designers to make informed decisions at the early design stage to decrease the construction of energy-inefficient buildings. This chapter ends with the thesis structure and the next chapter will review the literature on building energy consumption prediction and other appropriate areas pertinent to this research study.

CHAPTER TWO

2.0 OVERVIEW OF BUILDING ENERGY CONSUMPTION PREDICTION: METHODS, SIMULATIONS, TECHNIQUES, THEORY AND PERFORMANCE FACTORS

2.1 Chapter Overview

This chapter explores the fundamentals of building energy consumption, the traditional and contemporary method of building energy assessment at the design stage. This chapter also discusses the benefits and limitations of the traditional method and their impact on the construction process. Furthermore, this chapter explores the integration of Artificial Intelligence (AI) or Machine Learning (ML) tools for building energy consumption prediction. These contemporary methods offer great potential for enhancing the performance for energy consumption prediction and identifying energy-saving opportunities. This chapter highlights prominent AI/ML algorithms utilised for building energy consumption prediction. Additionally, this chapter highlights the key factors that affect the performance of these models in building energy consumption prediction. Understanding the challenges and limitations of these tool is fundamental for making informed decisions towards the development of an accurate and high performing models. Lastly, this chapter explores the theory of current practice.

2.2 Building Energy Consumption

Buildings engender a significant fraction of global energy consumption and carbon dioxide (CO₂) emissions worldwide. Yet, there is constant demand for more buildings as it is attributed to the increase in population and standard of living (Qiao et al., 2021). The population in the United Kingdom (UK) increased exponentially by 4.56 million from 2009 to 2019, such increase last occurred over a 35 years period between 1965 to 2000 (The World Bank, 2019). Based on the continuous growth, the UK population is estimated to increase by 5.17 million within the next 20 years (Office for National Statistics, 2019). Currently, the UK is also experiencing housing shortage on record with a deficit of four million houses (Barton and Wilson, 2021; Bulman, 2018). Thus, the UK government aims to build 300,000 new homes

every year until 2031 to equal demand (Bulman, 2018). The achievement of this set goal will fundamentally increase the amount of total energy consumption in the UK.

This increase has engendered several projections, for instance, the United Nations Framework Convention on Climate Change (UNFCCC) and United Nations Environment Programme (UNEP) projects that, total Green House Gas (GHG) emissions will double in the next 20 years. This would annihilate the possibility of reforming the construction sector and maintaining the Paris arrangement of improving energy intensity by 30% in 2030 (United Nations Environment Programme, 2017; United Nations Framework Convention on Climate Change, 2015). Likewise, heating and cooling energy consumption is projected to increase further, and this will be a result of the recurrent heat waves and cold spells due to global warming related climate change (Kadir Amasyali and El-Gohary, 2021). In 2017, the United Nations Environment Program (UNEP), reports that residential and commercial buildings employ about 60% of worldwide electricity, 40% of global energy, 40% of global resources, and emit approximately 1/3 of Green House Gas (GHG) (United Nations Environment Programme, 2017) and it is estimated that global energy consumption will grow by over 50% before 2030 (Mawson and Hughes, 2020). In the United Kingdom (UK), buildings accounted for over 29% of the total energy consumption in the UK (“BEIS,” 2019). Generally, the adverse effects of excess energy consumption include air pollution mainly in the form of CO₂ emission and its associated health effects [e.g. kidney disease, lung cancer, heart disease, among others (World Health Organisation, 2019)], climate change and global warming (Dandotiya, 2020), which deploys a severe impact on the existence of mankind (Dandotiya, 2020). According to the World Health Organisation (WHO), air pollution is one of the major causes of various adverse health problems on the population, as it intensifies the possibility of lung cancer, kidney disease, heart disease, among others (World Health Organisation, 2019). In the UK, air pollution is considered the most significant environmental threat to health, with over 28,000 deaths recorded yearly (Public Health England, 2019).

The global increase in energy consumption and the associated adverse effects trigger the importance of understanding building energy efficiency and exploring various potentials for enhancement (TOPRAK et al., 2017). This point is well understood globally as various governments are constantly executing regulations, principles and sometimes incentives that are tailored towards enhancing building energy savings initiatives (Amasyali and El-Gohary, 2021; Himeur et al., 2020; Qiao, Yunusa-Kaltungo and Edwards, 2021). For instance, In 2002, the

European Union (EU) mandate on building energy performance implemented a systematic structure for understanding energy efficiency which has since prompted member states to generate certification systems for rating building energy performance (European Parliament, 2002). Various institutions such as Building Research Establishment (BRE), Chartered Institute of Building Services Engineers (CIBSE), and Building Services Research and Information Association (BSRIA) have also made significant contributions to building services field, especially in the areas of sustainability, energy efficiency, and indoor environmental quality(BRE, 2024; BSRIA, 2024; CIBSE, 2024). BRE's publications focus profoundly on the advancement of standards and models for sustainable building design, including the BRE Environmental Assessment Method (BREEAM) and BRE Domestic Energy Model (BREDEM), which has been effective in the formation green building certification practices (Henderson and Shorrock, 1986; BRE, 2024). This method not only stimulates energy-efficient designs but also supports wider sustainability objectives such as resource conservation and waste reduction, aligning with current environmental goals in building services. CIBSE and BSRIA have produced widely recognized technical guidelines that address practical needs in building design and operation, especially concerning HVAC systems, lighting, and indoor air quality (Milivojevic and Ahmed, 2018; Bleicher, 2023; Marrow, 2023). Therefore, the UK employs a standard scale rating system to notify building owners of current energy performance and effective methods to improve energy efficiency (Curtis et al., 2014). In the industrial sector, the rise in energy demand accompanied by increasing energy prices have propelled companies to implement energy management strategies for improved use of energy and decrease in energy expenses. However, the execution of such strategies does require knowledge of both past and future energy demands (Mawson and Hughes, 2020).

Although most government policies are tailored more to the current building stock, research has emphasized the importance of considering energy efficiency at the design stage, to ensure future buildings are more energy efficient (Ding et al., 2018; Jaber and Ajib, 2011; Wang et al., 2014). It is stipulated that decisions made at the design stage of buildings can considerably impact the energy performance of a building, as design stage is where the potential lies for around 30% energy savings (Aversa et al., 2016; Colmenar-Santos et al., 2013). One key method to reduce energy consumption and ameliorate the construction of energy-inefficient buildings is by enabling energy modelling for designers during the design stage. Currently, designers use building energy simulation tools at this stage, which demand numerous parameters and substantial processing time (from a few hours to several days) to evaluate

potential energy performance(Jhai, 2023; Pham et al., 2020). Sometimes, designers optimize designs based on past experiences, often leading to inaccurate results (Deb et al., 2017a; Runge and Zmeureanu, 2019; Wang et al., 2005). Accurate energy prediction models could potentially save about 30% of a building's total energy use (Aversa et al., 2016; Colmenar-Santos et al., 2013). Recognizing the benefits of predicting building energy consumption and the necessity to improve energy efficiency understanding, researchers have focused on this area, leading to new advancements through various techniques(Ahmad et al., 2017a; Fathi et al., 2020a; Serale et al., 2020; Bourdeau et al., 2019; Runge and Zmeureanu, 2019; Amasyali and El-Gohary, 2018). There are two major techniques for energy modelling and prediction are physical (Chirarattananon and Taveekun, 2004; Yezioro, Dong and Leite, 2008; Neto and Fiorelli, 2008) and data-driven method (Li et al., 2009a; Niu, Wang and Wu, 2010; Ahmad, Mourshed and Rezgui, 2017; Wang et al., 2018).

2.3 Energy Simulation Tools (Physical Method)

The Physical method is also known as the physics-based modelling approach, is the utilization of simulation tools for analysing and predicting building energy consumption (Qiao et al., 2020a). They depend on thermodynamic rules for thorough energy analysis and modelling (Amasyali and El-Gohary, 2018). The simulation tools based on this approach include EnergyPlus (Neto and Fiorelli, 2008), DOE-2 (Chirarattananon and Taveekun, 2004) and eQuest (Yezioro et al., 2008), Ansys among others. EnergyPlus and Ansys enables users to specify energy values and generate energy-efficient designs and they are widely used for building performance simulations to optimize energy usage(Neto and Fiorelli, 2008). The parameters required for physical modelling such as insulation thickness, thermal properties, building systems need to be acquired from physical features, often from design plan and on-site measurement. This method is recognized as the traditional method utilized at the design stage for building energy assessment. It enables designers to upload building designs to access the potential energy performance of the building. Figure 2.1 summarizes the data flow and key processes of the physical model development and simulation.

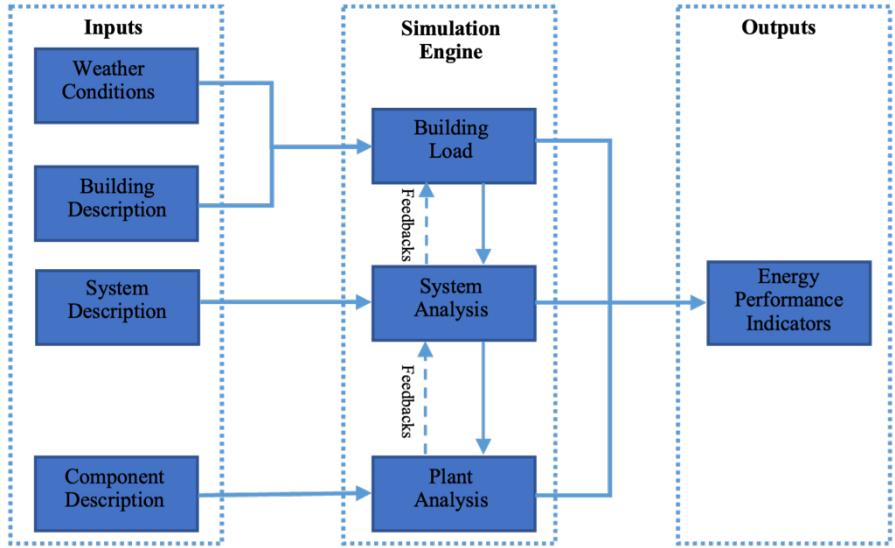


Figure 2.1: Data flow and key processes of physical model simulation (Li and Wen, 2014)

For several years, energy simulation tools such as EnergyPlus and eQuest have been applied for assessing energy consumption however, it elicited some advantages and disadvantages (Crawley et al., 2008). The advantage of using simulation tools is the clear correlation between input and output parameters. However, considering the drawbacks, simulation tools requires a large number of building parameters that are often inaccessible (e.g., HVAC (Heating, Ventilation and Air Conditioning) system, internal occupancy loads, physical properties and solar information among others) (Deb et al., 2017a; Li and Wen, 2014; Qiao et al., 2021; Runge and Zmeureanu, 2019) and consumes a lot of time which often hinders designers work flow (Pham et al., 2020; Zhu, 2006). Studies such as Al-Homoud, 2001 provided a very comprehensive study on energy simulation tools. However, research on energy simulation tools has significantly declined, subject to its stringent requirements of excess building and environmental features, which are not often readily available at the time of simulation, resulting in poor building energy prediction performance (Hai-xiang Zhao and Magoulès, 2012a). Many factors influence building energy behaviour such as meteorological conditions, occupancy behaviour, complex interactions of heating, ventilation and air conditioning (HVAC) etc which make it difficult for computer-based energy simulation tools to achieve accurate calculations of energy consumption (Deb et al., 2017b; Mocanu et al., 2016).

According to Ahmad et al., (2017), simulation tools do not produce good performance in predicting building energy use and they are rendered unsuitable for real-time implementation, as the intricacy and difficulty of utilizing this method will rise when applied to multiple buildings (Ahmad et al., 2017a; Pham et al., 2020). Zhu further emphasized that, although computer-based simulation tools are considered beneficial to faculty managers for detecting

better energy conservation solutions and providing designers with relevant insights towards reducing energy consumption, the model creation process is time consuming and resource demanding (Zhu, 2006).

Conversely, Data driven method is often developed using mathematical models or Machine Learning (ML) algorithms [e.g. Support Vector Machine (SVM) (Li et al., 2009a; Niu et al., 2010) and Artificial Neural Networks (ANN) (Ahmad et al., 2017a; Runge and Zmeureanu, 2019) among others], which require less number of building parameters. Research has proffered data driven model as a more efficient approach for predicting energy consumption due to good performance, lesser requirements and time efficacy (Runge & Zmeureanu, 2019; Qiao, Yunusa-Kaltungo and Edwards, 2020; Neto and Fiorelli, 2008; Yezioro, Dong and Leite, 2008). For example, Neto & Fiorelli (2008) conducted a comparison between a data-driven model and a physical model (EnergyPlus) for predicting building energy consumption. It was concluded that the data driven model indicated better performance than the physical model (Neto and Fiorelli, 2008). In 2008, Yezioro et al. applied data driven model and physical model (Energy_10, EnergyPlus, eQuest) for estimating building energy performance. The performance evaluation reveals a good prediction with the data driven model as against the simulation models (Yezioro et al., 2008).

As stated above, several studies have conducted comparative analysis between data driven and physical methods, which conveys that data driven method outperforms the physical methods in building energy prediction (Ahmad et al., 2017a; Chirattananon and Taveekun, 2004; Neto and Fiorelli, 2008; Yezioro et al., 2008; Zhu, 2006). Hence this research only focuses on data driven method because despite the good properties of data driven method, data driven method possess several challenges as can be seen in the next section.

2.4 Artificial Intelligence-based Tools (Data Driven Method)

In recent times, data driven method has posed new prospects for building energy use prediction (Qiao, Yunusa-Kaltungo and Edwards, 2020). Data driven models are centred strictly on the utilization of mathematical models or Machine Learning (ML) algorithms [e.g. Support Vector Machine (SVM) (Li et al., 2009a; Niu et al., 2010) and Artificial Neural Networks (ANN) (Ahmad et al., 2017a; Runge and Zmeureanu, 2019) and Random Forest (RF) (Ahmad et al., 2017a; Z. Wang et al., 2018b) among others] which do not require much detailed information of building (Runge and Zmeureanu, 2019).

Machine learning (ML), a subfield of Artificial intelligence (AI) and by definition, AI is a wide field encompassing computer systems that possess the ability to implement a task that customarily require human intelligence, for example visual perception, speech recognition and decision making under uncertainty (Russell and Norvig, 2020). Meanwhile, ML is defined as the study of algorithms and statistical models that computer systems use to perform a certain task without being explicitly programmed (Yu, 2020). ML can be used to solve various problems, such as spam identification, customer segmentation detecting fraudulent transactions, image and video recognition and making product recommendations based on historical information.

Machine Learning (ML) method is considered one of the most suitable method for prediction tasks, often producing desired outcomes (Ríos Canales, 2016; Vorobeychik and Wallrabenstein, 2013). In the past decade, several machine learning algorithms such as Artificial neural network (ANN), Deep Neural Network (DNN), Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF), Decision Tree (DT), Stacking, K Nearest Neighbors (KNN), Extreme Gradient Boosting Trees (XGB) have been applied in the field of building energy use prediction (Aversa et al., 2016; Chou and Bui, 2014; Dong et al., 2021a; Cheng Fan et al., 2017; K. Li et al., 2018; Li et al., 2009b; Pham et al., 2020; Tardioli et al., 2015; Wang et al., 2020). These algorithms have proven suitable for different building groups [residential building (Curtis et al., 2014), non-residential (Chirarattananon and Taveekun, 2004)], other prediction time period [hourly, monthly and annually (Dong et al., 2021a; Qiong Li et al., 2010)] and different prediction purposes [heating or cooling load, total consumption (Li et al., 2009b)]. Despite the good performance of data driven method using ML algorithms, past studies have not been extensively developed energy prediction models for design stage analysis which is where the potential lies for around 30% energy savings (Aversa et al., 2016; Colmenar-Santos et al., 2013). Hence, there are not many studies to substantiate its suitability of ML algorithms for prediction at the early design stage of buildings.

ML algorithms are becoming more recognised due to their low time consumption and good performance in building energy use prediction (Seyedzadeh et al., 2020). However, the good performance of ML algorithms is not limited to the energy prediction world, even in other subject areas such as spam filter, cancer prediction, among others. For instance, Dada et al., (2019) applied some machine learning algorithms (i.e. Support Vector Machine (SVM), Naïve Bayes, K Nearest Neighbour (KNN), Neural Network, among others) in the field of email spam

filtering and concluded that machine learning classifiers produced good result in classifying spam email. It was further recommended that future study should explore the use other ML algorithms, deep learning and deep adversarial learning algorithms (Dada et al., 2019). Likewise, in the world cancer prediction, Al-Shargabi and Al-Shami, (2019) conducted an experimental study using three ML algorithms (Random Forest (RF), K Nearest Neighbour (KNN) and Multilayer Perceptron (MLP)) for breast cancer predictions. The best algorithm was selected in terms of accuracy. Furthermore, KNN and RF produced more accurate results than MLP (Al-Shargabi and Al-Shami, 2019).

In the energy prediction world, several studies have proposed the utilization of machine learning algorithms for building energy prediction (Ahmad et al., 2017a; K. Li et al., 2018; Pham et al., 2020; Qiong Li et al., 2010; Caleb Robinson et al., 2017b). A few of such algorithms have proven to be effective for predictions such as SVM, and ANN. Although ML algorithms produce good performance in various subject areas, their prediction performance is reliant on these three factors: algorithm selected, quantity and quality of the data (Lee et al., 2011; Kaur and Gupta, 2017; Runge and Zmeureanu, 2019; Dalal, 2018; Goyal, Tiwari and Sonekar, 2020; Kabir, 2020).

2.4.1 Machine Learning Algorithms

An “algorithm” in machine learning is a computational technique that can learn from sample data and predict output values based on input data (Abdollahi et al., 2019). There are many popular algorithms that have evolved over the years.

2.4.1.1 Types of Machine Learning Algorithms

The different types of Machine Learning algorithm include Artificial Neural Network (ANN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF) among others.

2.4.1.1.1 Artificial Neural Network (ANN)

Artificial Neural Networks are the most broadly utilised for predicting building energy consumption (Ahmad et al., 2014; Bourdeau et al., 2019). ANN is a non-linear computational model that emulates the functional concepts of the human brain (Amasyali and El-Gohary, 2018). ANN is an effective approach for solving non-linear problems and is dominant with big datasets which enable the neural network sufficient data to train the model (Bourhnane et al., 2020). There are several types ANNs such as Back Propagation Neural Network (BPNN), Feed

Forward Neural Network (FFNN), Adaptive Network-based Fuzzy Inference System (ANFIS) etc. Among them, feed-forward is the most frequently utilised (Ahmad et al., 2017a). Multi-layer Perceptron (MLP) is a function of deep neural network that utilizes a feed forward propagation process with one hidden layer where latent and abstract features are learned (Donoghue and Roantree, 2015)

The basic form of ANN consists of three consecutive layers namely input, hidden, and output layer as illustrated in Figure 2.2 below. The input layer is used for train the model, the hidden layer is the bridge between input and output layer which can be modified dependent on the type of ANN while the output layer provides the result (Bourdeau et al., 2019). There are several types ANNs such as Back Propagation Neural Network (BPNN), Feed Forward Neural Network (FFNN), Adaptive Network-based Fuzzy Inference System (ANFIS) etc. Among them, feed-forward is the most frequently utilised (Ahmad et al., 2017a). Figure 2.2 displays an illustrative diagram of a feed forward neural network architecture, containing of two hidden layers.

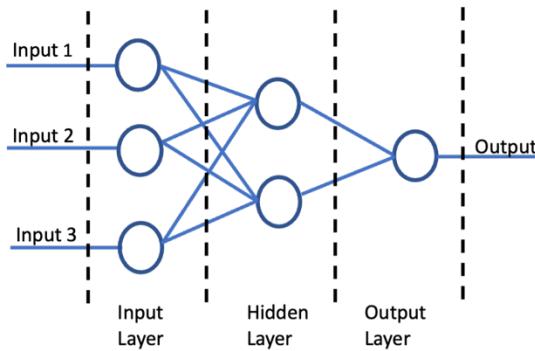


Figure 2.2: Feed Forward Neural Network Architecture.

2.4.1.1.2 Support Vector Machine (SVM)

SVM is a data mining algorithm, recognised as one of the most robust and accurate methods between all data mining algorithm (Wu et al., 2008). SVM is increasingly used in research due to its ability to effectively provide solutions to non-linear problems in various sizes of data (Hai-xiang Zhao and Magoulès, 2012b). SVM has gained more attention, owing to its capability of effectively generating good solutions to non-linear problems in diverse sample sizes (Chen et al., 2022; Olu-Ajayi et al., 2023a; Hai-xiang Zhao and Magoulès, 2012b). SVM is based on the kernel, a method primarily computed for solving binary classification problems proposed by Vapnik in the early 1990s (Sonkamble and Doye, 2008). The SVM utilized for regression is known as Support Vector Regression (SVR), which has emerged a significant data driven method for forecasting building energy use. The main task in SVR is to create a

decision function, $F(x_i)$, by using a historical data also called the training process. For the given input x_i , it is required that the outcome predicted should not differ from the real target y_i greater than the predetermined threshold ε . This function is often expected in form of

$$F(x_i) = (\varepsilon, \varphi(x_i)) + b \quad (1)$$

It is worth highlighting that SVM, or more specifically SVR are superior to other models because its framework is effortlessly generalized for various issues and it can achieve optimum solutions worldwide (Wei et al., 2018).

2.4.1.1.3 Random Forest (RF)

Random forest is an ensemble technique that offers several beneficial features. Among these features are (Ahmad et al., 2017a); (i) It is built on ensemble learning theory, which enables it to learn both simple and convoluted problems. (ii) It does not demand much hyper-parameter tuning to achieve good performance in comparison to other ML algorithms (e.g, artificial neural network, support vector machine, etc.). (iii) It's default parameters often produce excellent performance. Hence, the Random Forest (RF) method is gaining more attention in the field of building energy consumption (Ahmad et al., 2017a; Y.-T. Chen et al., 2019; Cheng Fan et al., 2017; Z. Wang et al., 2018b). For example, Pham et al., (2020) conducted the application of Random Forest (RF), model tree and Random Tree (RT) for forecasting hourly energy load. It was concluded that Random Forest (RF) produced the most suitable result.

2.4.1.1.4 Gradient Boosting (GB)

Gradient Boosting algorithm is a machine learning method that can be utilized for both regressor and classification problems. This technique algorithm builds model in stages like other boosting systems but generalizes these by enhancing an arbitrary differentiable loss function (Flores and Keith, 2019). The Gradient Boosting method utilizes an ensemble of weak models which collectively form a stronger model. The final model is a function that receives a vector of attributes $x \in R^n$ as input to identify a value $F(X) \in R$. Furthermore, one of the reasons for utilizing GB is based on the preceding reputation of ensemble methods outperforming other machine learning techniques in various situations (Berk, 2006; Dietterich, 2000; Zhang and Zhang, 2009). They are generally recognized as the regressors or classifiers that produce the best out-of-the-box results. GBM and RF have similar training processes, except for a primary difference, RF trains each decision tree independently, by utilizing random parameters, and merges the outcomes from all independent tree, while GBM trains decision

trees one after the other, with the new trees bidding to reduce errors created by previous trees, a method recognised as boosting.

2.4.1.1.5 Decision Trees

Decision Tree (DT) is a method of utilizing a tree-like flowchart to partition data into groups. Decision Trees is an adaptable process that could advance with an enlarged amount of training data (Domingos, 2012). In contrast to other data-driven methods, DT is easier to comprehend, and its application does not require complex computation knowledge. However, it often produces major deviation of its predictions from actual results. DT is more suitable for forecasting categorical features than for estimating numerical variables (Yu et al., 2010).

DT commences execution at the root node where input data are split into various groups dependent on some predictor variables pre-set as splitting criteria. These split data are then dispersed into the branches originated from the root node denoted as sub-nodes. These sub-nodes data will either undergo further or no splits. The internal data node is where further data split is conducted to develop new subgroups. However, the concluding are the leaf nodes that handles the corresponding data groups at the current level as their final outputs. Figure 2.3 is an example DT representation utilised for medium annual source energy use per unit floor ($\text{kWh/m}^2/\text{yr}$) of a non-residential building. In which, the building consumption ratio and gross floor area are selected as predictor variables in the root and internal node respectively, adopted from Wei et al., (2018).

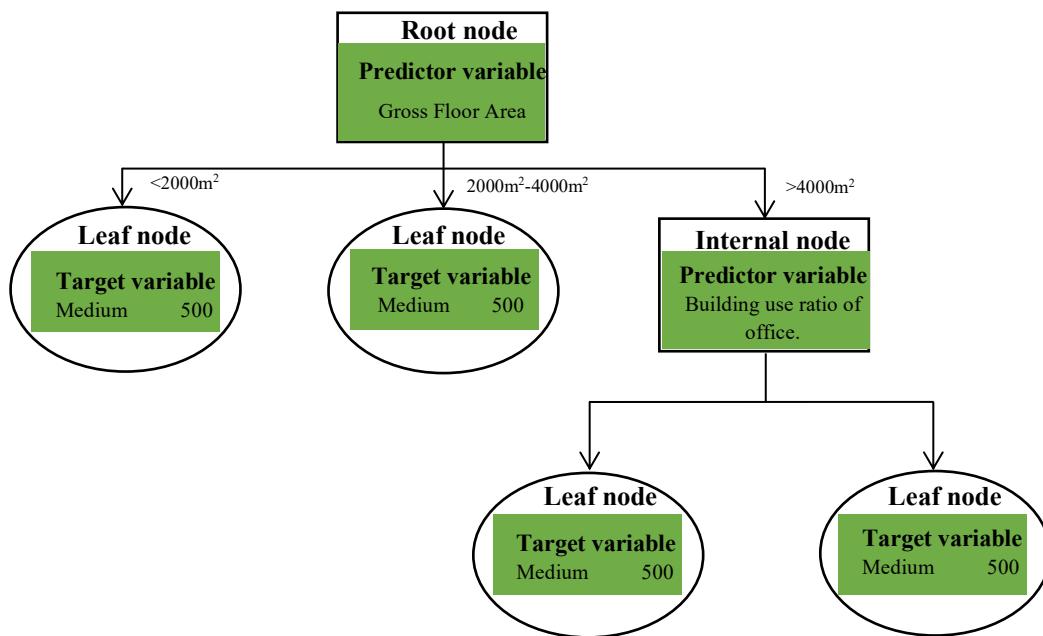


Figure 2.3: Illustrative Diagram of a medium annual energy use per unit floor

2.4.1.1.6 K-Nearest Neighbour (kNN)

K-Nearest Neighbour (k-NN) algorithm is a non-parametric machine learning method that utilizes similarity or distance function d to predict outcomes based on the k nearest training examples in the feature space (Ortiz-Bejar et al., 2018). kNN algorithm is one of the common distance functions that works effectively on numerical data (Ali et al., 2019). However, KNN is yet to receive much attention in the field of building energy prediction. In the study by Feng et al., KNN produced good results in predicting building energy use with an R^2 of 0.84 (Wang et al., 2020). The prediction for KNN as a regressor is performed as follows: From an input of x_i and output y_i is deduced based on the nearest record or most similar (nearest neighbor) $x \in X$. Figure 2.4 illustrates the procedure for selecting the nearest neighbour in values of ϵ . For example, squares labelled 1 and 2 will be chosen on query with radius ϵ_2 , contrarily, no squares will be retrieved on a query with radius ϵ_1 . The radius must be increased until one element is selected (Olu-Ajayi, 2017).

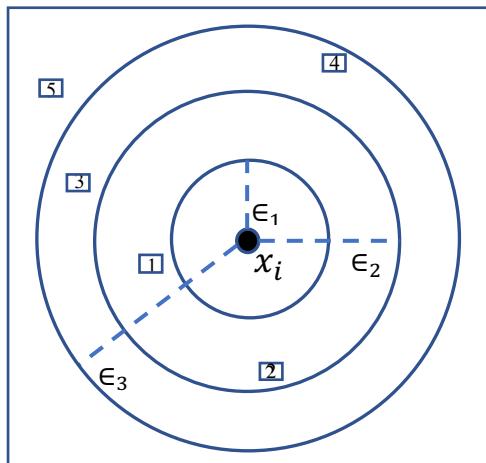


Figure 2.4: Example of k-NN Regressor

2.4.1.1.5 Deep Neural Network

In recent years, the deep learning methods are providing powerful techniques to achieve enhanced modelling and better prediction performance (C. Li et al., 2017). The deep learning method uses deep architectures or multilayer architectures. It is a basic structure of deep neural networks, and the main distinction between Deep Neural Networks (DNN) and shallow neural networks is the number of layers. Generally, shallow neural networks have only two to three layers of neural networks which limits its ability to express intricate functions (Lei et al., 2021). Conversely, deep learning has five or more layers of neural networks and presents more efficient algorithms that can further increase the accuracy.

Deep Neural Networks method is considered a superior ML technique due to the addition of multiple hidden layers to the regular ML neural network and it has gained increased attention in various fields such as image recognition (Yu et al., 2021) and natural language processing (Leyh-Bannurah et al., 2018) among others. However, Deep Neural Networks (DNN) have not received much attention in the field of building energy consumption prediction (Kadir Amasyali and El-Gohary, 2021). Figure 2.5 visualizes the deep neural network architecture.

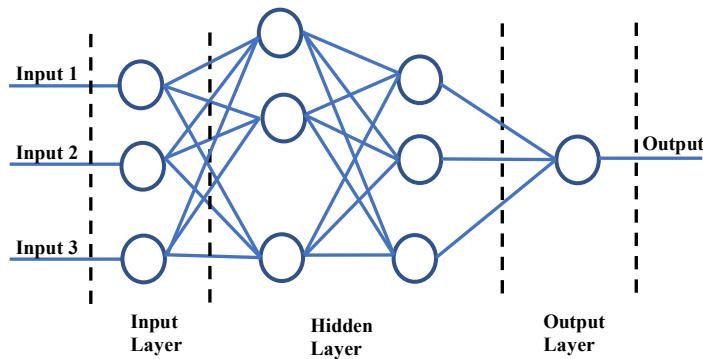


Figure 2.5: Deep Neural Network Architecture

The advancement of ML method has produced satisfactory results in energy prediction (C. Li et al., 2017). Support Vector Machine (SVM) is one of the recurrent algorithms in the field of energy prediction and it is known for its delivery of good results in small datasets (Aversa et al., 2016; Li et al., 2009b; Qiong Li et al., 2010). The application of SVM for building energy consumption predicting was first proposed by Dong et al, and it was determined that SVM produced better performance than other related research using neural networks (Dong, Cao and Lee, 2005). Dong et al (2016) also applied data driven model using machine learning algorithms, specifically Support Vector Regression (SVR), Artificial Neural Network (ANN) and hybrid methods namely Least-Square Support Vector Machine (LS-SVM) and Gaussian Process Regression (GPR) to predict electricity usage. It was determined that the hybrid method produces better results for hourly energy predictions (Dong et al., 2016).

Wang et al (2018) applied Ensemble Bagging Tree (EBT) to predict the energy consumption using a commercial building (Z. Wang et al., 2018a). Ensemble models comprises of various algorithms and due to their stability, they are likely to generate better outcomes than single models (Kadir Amasyali and El-Gohary, 2021). These types of ensemble model encompass Random Forest (RF), Gradient Boosting (GB), Ensemble Bagging Tree (EBT), among others. Furthermore, in research by Chae et al (2016), ensemble models were compared to other ANN

algorithms to predict building energy consumption. The outcome showed the ensemble models outperforms ANN algorithms (Chae et al., 2016). In Hong Kong, Tso and Yau (2007) explored and compared the utilization of neural networks, decision tree and regression methods for weekly electricity consumption prediction. Result shows that decision tree and neural networks achieved slightly better results than the regression technique with a root of average squared error (RASE) of 39.36 (Tso and Yau, 2007). Although, SVM and ANN have been noted as most suitable in past studies, they have at different times outperformed one another. For instance, the study by Dong et al (2005), concluded that SVM produced better results than neural networks (Dong, Cao and Lee, 2005); while Khantach et al (2019) conducted a comparative analysis of Multi-layer Perceptron ANN and SVM among others. It was concluded that ANN outperformed SVM for energy use prediction (Khartach et al., 2019).

It is evident the investigation of the best algorithm remains a complex task, as there is no general agreement on the most suitable algorithm to use for energy prediction (Amasyali and El-Gohary, 2017). Additionally, no single algorithm can be selected as the most suitable, unless the model is developed and compared on the same quantity and quality of data (Zhong et al., 2019).

2.4.1.2 Quantity of the Data

In the machine learning world, it is a general hypothesis that the larger the data used to train the model, the better the performance and more reliable the result (Dalal, 2018; Goyal et al., 2020; Kabir, 2020; Kaur and Gupta, 2017; Lee et al., 2011). Several studies have explored the comparison of various ML algorithms using different quantity of data (Dong et al., 2021a; Cheng Fan et al., 2017; Pham et al., 2020; Wang et al., 2020). For example, In 2019, Runge and Zmeureanu examined the application of ANN for predicting hourly building energy consumption and further established that ANN algorithm was implemented using a single commercial building dataset and produced poor hourly predictions (Runge and Zmeureanu, 2019), which could suggest that the poor performance achieved is subject to the single dataset used to train the model. Bagnasco et al (2015) applied ANN for electrical consumption prediction using data from a single hospital building centred on weather-related data and time/day difference. ANN prediction achieved better results during winter (Bagnasco et al., 2015). Li et al (2009) conducted an application of SVM for predicting hourly cooling load using a single office building. It was established that SVM yields a good result for predicting hourly load (Li et al., 2009). This good performance of SVM can be subjected to the hypothesis that SVM generates good result in small datasets (Aversa et al., 2016; Li et al., 2009b; Qiong

Li et al., 2010). Furthermore, the study by Dong et al (2021) compared the application of ANN and SVM for hourly energy consumption prediction using a dataset of 507 non-domestic buildings. The weather data (atmospheric pressure, outdoor temperature, wind speed etc.) and building data (floor area and building type) were utilized as input features. Dong et al specified that ANN outperformed SVM (Dong et al., 2021a).

Furthermore, Fan et al., (2017) compared a number of algorithms (Deep Neural Network (DNN), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting Machines (GBM) among others) for predicting short term cooling load using one non residential building. Extreme Gradient Boosting Trees (XGB) produced relatively good results for cooling load prediction. Additionally, Pham et al., (2020) explored the application of Random Forest (RF), model tree and Random Tree (RT) for predicting hourly load using 5 buildings datasets. Random Forest(RF) emerged the most efficient. Runge and Zmeureanu (2019) further highlighted that the development of an efficient energy use prediction model should be based on a large dataset (Runge and Zmeureanu, 2019).

Despite the good performance of ML algorithms in evident from past studies (e.g. Aversa et al., 2016, 2016; K. Li et al., 2018; Pham et al., 2020; Robinson et al., 2017), it is proffered that the utilization of even larger quantity of data will engender better performance (Runge and Zmeureanu, 2019).

2.4.1.3 Quality of the Data

In the machine learning world, Data pre-processing is essential for dataset quality assurance, however, it can be computationally expensive and time consuming (Shapi et al., 2021). The processing of data prevents complexities when training the model such as missing or abnormal values. Another method of improving the quality of a dataset is feature selection. Feature Selection (FS) aims to improve the performance of ML models by eliminating the unimportant and irrelevant noisy features, thus improving the quality of the dataset (Asir et al., 2016). These unrelated features that constitute no correlation to labels serve as pure noise, which could lead to bias in prediction, thereby diminishing the classification performance (Kunasekaran and Sugumaran, 2016). Such situations require feature selection to speed up the learning process and enhance the quality of data. Ahmad et al., (2017a) utilized feature selection in the development of Random Forest (RF) and Artificial Neural Networks (ANN) for the building energy use prediction and concluded that it produced good performance. Additionally, Faisal et al., (2019) applied feature selection to calculate the relevance of the input features for

predicting electricity consumption. It was concluded that the results produced using the selected features outperformed the results using the original features. Feature selection (FS) is discussed in further detail in chapter 6.

2.5 THEORY REVIEW

This section examines the key theory relevant to this research. The theory named forecasting theory is the foundational theoretical framework of this research and it is discussed below.

2.5.1 Forecasting Theory

The forecasting theory is solely based on the grounds that past and current knowledge can be utilised to make future predictions(Petropoulos et al., 2022). It is noted that forecasting techniques perform best when applied to solve a problem in practice. In understanding the relevant features of the problem, the theory can be deduced. Fundamentally the theoretical results can engender improved practice.

The utilization of big data have been proliferated in literature in the last two decades (Alaka et al., 2019; Balogun et al., 2021; Swanson and Xiong, 2018). However, despite the promised enhancement in forecast accuracy, there remains unsubstantiated evidence to ascertain the improvements in energy forecast. This research also investigates if big data engenders significantly increase the performance of energy forecasts. (Athey, 2019) contends the ability of machine learning techniques to be an efficient method handling with big data sets, and this is further investigated before establishing their inability to handling energy data. Feature selection aids the eradication of the unimportant and irrelevant noisy feature to further enhance the quality of the dataset (Asir et al., 2016). This research explores the effect of lower and more features of energy data has on the performance on statistical and machine learning techniques.

Spiliotis et al., (2020) proclaims that ML techniques are data-driven, thus they are more generic and easily adaptable to different types of forecasting. However, these methods encompass some limitations. It is stipulated that ML algorithms harnesses full capacity when sufficient data is employed in the development of the model. This research also conducts a reliability analysis to evaluate the effect of data on model performance.

Machine learning algorithms (see section 2.4.1) essentially lead to sub-optimal forecasts when data are not appropriately pre-processed (Makridakis et al., 2018). However, machine learning (ML) techniques can be successfully applied when working with long, high-frequency series,

which are typically found in applications related to the stock market (Moghaddam et al., 2016; Pawar and Jaiswal, 2020), and demand (Alduailij et al., 2021; Hribar et al., 2019).

2.6 Chapter Summary

This chapter provides a comprehensive examination of methods for predicting building energy consumption. It begins with an overview of the chapter, followed by an exploration of building energy consumption and the tools employed for prediction. The two major methods are discussed: physical methods using energy simulation tools and data-driven methods employing artificial intelligence. Within the data-driven approach, various machine learning algorithms are examined, including Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB) among others. Considerations of data quantity and quality are also addressed. The chapter concludes with a review of forecasting theory, laying the theoretical foundation for the predictive methodologies discussed.

CHAPTER THREE

3.0 SYSTEMATIC LITERATURE REVIEW AND THEORITICAL FRAMEWORK: STATISTICAL AND ARTIFICIAL INTELLIGENCE BASED TOOLS FOR BUILDING ENERGY CONSUMPTION PREDICTION

3.1 Chapter Overview

This chapter conducts a comprehensive and systematic review of literature to identify the most suitable statistical and AI based tools for predicting building energy consumption. This chapter systematically analysed journal articles that employed statistical and AI tools in building energy prediction (BEP). This chapter evaluates the performance of nine popular and promising statistical and AI tools with a primary focus on 7 pertinent criteria (e.g., data size, error rate etc.) in the building energy research domain. Furthermore, one of the key contributions of this chapter is the development of a diagrammatic framework, carefully curated to serve as a guide for appropriate tool selection in various situations in the field of building energy consumption prediction. Essentially, this will equip researchers and practitioners with a streamlined framework for tool selection, thereby saving time and effort consumed on exhaustive comparative analysis of several tools to identify the most suitable for a specific situation. Furthermore, this chapter conducts a quantitative bibliometric analysis to pinpoint the trends and examine knowledge gaps. Also, this chapter helps to achieve Objective #2 of this research.

3.2 Significance of Tool Evaluation

The explosive rate of energy consumption and its ecological impacts are considered major challenges all around the world, as the unparalleled increase in the global population alongside economic advancements, comfort index, industrialization, among others have an eminent effect on global energy consumption(Deb et al., 2017b; Somu et al., 2020). The primary solution to these challenges is energy conservation, as the United Nations (UN) population division projects a significant increase from 7 billion to 8.3 billion by the year 2030 (United Nations Population Division, 2017). Yet, there is a limited volume of natural resources for providing the required energy for humans to conduct their day-to-day lives (Fathi et al., 2020b). Many countries are experiencing energy problems in varying intensities, for example, energy

consumption for residential space heating and cooling accounts for 70% in the United Kingdom(UK) (Li and Yao, 2020), 47% in the united states (US) (H. Chen et al., 2021), and 64% and 69.58% in the European Union (EU) and China, respectively (C. Liu et al., 2020; Shakya et al., 2021). The building sector has emerged as the most prominent energy consumer, as humans spend 90% of their daily lives in buildings, accounting for over 85% of total energy consumption of buildings life cycle, thereby contributing over 39% and 38% to global energy consumption and carbon dioxide (CO₂) emission respectively (Chammas et al., 2019; Sadeghi et al., 2020; Somu et al., 2020).

The deficiency of energy resources, continuous increase in energy demand and environmental degradation have engendered an increase in research focused on energy efficiency and the reason for this increase is undoubtedly justified (Alduailij et al., 2021; K. Amasyali and El-Gohary, 2021; Carrera et al., 2021; Olu-Ajayi et al., 2022a). According to the International Energy Agency, building energy efficiency is an imperative measure for global sustainability and long term decarbonization (“Climate change,” 2011; Sha et al., 2019). Given these impacts, several methods are continuously researched to improve energy efficiency ubiquitously (Badiei et al., 2020; Bourdeau et al., 2019; Cajias and Piazolo, 2012; Fan et al., 2020; Li et al., 2020). Of these methods, building energy consumption prediction tools has attracted more attention from researchers globally (Carrera et al., 2021; Chokwitthaya et al., 2020; Dun and Wu, 2020; Y. Liu et al., 2020a; Sadeghi et al., 2020). It is considered the most promising and potentially efficient solution for achieving energy efficiency, energy waste reduction and alleviating climate changes (Khan et al., 2021; Olu-Ajayi et al., 2022b; G. Zhang et al., 2020). As noted in section 2.4, The performance of a prediction models is highly predicated on the tool selected, among other factors (Goyal et al., 2020; Kabir, 2020; Runge and Zmeureanu, 2019). Nonetheless, excluding a few studies (e.g. Culaba et al., 2020; Shao et al., 2020), tool selection in numerous building energy prediction (BEP) studies is centred more on popularity (Divina et al., 2018a; C. Fan et al., 2017; Feng and Zhang, 2020; Somu et al., 2021); than its capabilities. This can be attributed to the absence of many tool evaluation studies that examine the comparative performance of the major tools based on several relevant criteria a BEP model should fulfil.

In the past few decades, the most employed tools for BEP are the statistical or artificial intelligence (AI) tools (Amber et al., 2018; Chou and Tran, 2018). An AI-based tool known as Artificial Neural Network (ANN or NN), was identified as the most common for energy

consumption prediction (Ahmad et al., 2017b; J.-H. Kim et al., 2020). It is designed to emulate the learning processes of the human brain (Ngarambe et al., 2020). Due to its popularity, several studies arbitrarily applied feed forward and back propagation ANN tool for BEP (Almalaq and Zhang, 2019; Hwang et al., 2020; Izidio et al., 2021; Koukaras et al., 2021; Lee and Rhee, 2021; Shan et al., 2019) among others. Despite its reproval which includes moderately adverse features such as computational intensity, and lack of transparency (Alaka et al., 2018; Mat Daut et al., 2017; Sadeghi et al., 2020), several studies have conducted a comparative analysis of ANN and other relatively popular tools (such as Random Forest (RF), Gradient Boosting Machine (GB or GBM)) on small sample size (Bouktif et al., 2020; Gao et al., 2020; Groß et al., 2021) disregarding the known theory that ANN requires large samples to generate optimal outcome (Alaka et al., 2018; Bourhnane et al., 2020; Olu-Ajayi et al., 2022b). Contrarily, of all AI-based tools, Support Vector Machine (SVM) has recently become the most common tool due to its ability to produce the optimum results for non-linear problems using small sample sizes (Aversa et al., 2016; Mat Daut et al., 2017; Olu-Ajayi et al., 2021). Although SVM is considered a powerful tool, its main weaknesses include large requirements of memory and training time consumption (Zhang et al., 2005). SVM was first employed by Dong et al., (2005) to develop an energy predictive model for monthly building energy consumption. Based on the comparison, Dong (Dong et al., 2005) stipulated that SVM outperform NN. Nonetheless, based on the popularity of these tools, recent research has employed both tools for comparison where NN outperformed SVM and vice versa (Chammas et al., 2019; Feng and Zhang, 2020; Li and Yao, 2020). The application of prediction tools should not be justified based on the acknowledgement of their strengths and recognition but also the acknowledgement of their limitation as well (Chung et al., 2008).

Considering there is no one size fits all AI-based tool, comparison and evaluation of these tools is paramount to avail BEP model developers a guideline towards an informed selection of tools (K. Amasyali and El-Gohary, 2021). A model developer should recognise the benefits and drawbacks of the existing tools to achieve good outcomes in relation to certain criteria (e.g., data size, error rate etc.). This will ascertain the appropriate tool is applied in the appropriate situation for the appropriate data features and purpose. Notwithstanding the importance of previous studies (Ahmad et al., 2014; Kuster et al., 2017; Seyedzadeh et al., 2018), there is still a deficit in review studies that evaluate existing AI-based tools for BEP from a more multivariate approach; encompassing the data side i.e. the size of data, and data type among others. Thus, by means of a structured and comprehensive review of studies that employed AI-

based tools, this research will produce a simplified framework for the selection of building energy consumption prediction tools in relation to relevant criteria.

This review is however limited to the review of prominent tools that have been applied in the development of energy consumption prediction models in previous studies. This is due to the impracticability of conducting a comprehensive review of all the numerous tools that can be employed for model development. The process of selecting the popular and favourable tools involved the review of conference and journal papers, among other academic publications, and also conducted a citation analysis of these papers. The outline of this review is structured as follows: this section is followed by a discussion of the other recent systematic literature review studies and their objectives, the next section explains the methodology for the systematic literature review conducted in this research, comprising of the inclusion and exclusion criteria and presents the result of bibliometric analysis. Section three follows with a brief explanation of the nine tools reviewed in this research; Section four describes the identified criteria employed in tool assessment. Section five conveys the result and findings of the systematic literature review; section six describes the proposed model, the theoretical and practical implication of this review, while section seven provides the conclusion drawn from this research.

3.3 Existing Review Studies on Building Energy Prediction Tools

In recent years, there has been a considerable increase in research focused on data-driven tools for building energy consumption prediction (Amasyali and El-Gohary, 2018). In response, several studies have conducted a systematic review of the existing data-driven tools. For example, Seyedzadeh et al., (2018) comparatively reviewed four data-driven methods namely SVM, ANN, clustering and Gaussian-based regressions that have been popularly employed for the prediction and enhancement of building energy performance. This research highlighted and focused on the strengths and drawbacks of these tools and their applications. Mat Daut et al., (2017) conducted a review and analysis of Stochastic time series methods and Hybrid ANN and SVM solely focused on the performance of these methods for electrical BEP. Kuster et al., (2017) focused on the review of studies with the application of time series analysis tools (Autoregressive Integrated Moving Average (ARIMA) and the Autoregressive Moving Average (ARMA)) and AI-based tools ANN and SVM for the prediction of electric consumption in buildings. It was proffered that the direct comparison of models across studies is futile. Furthermore, Wang and Srinivasan, (2017) presented a review of AI-based BEP tools

namely multiple linear regression (MLR) and ANN, among others with a key focus on ensemble models. Yildiz et al., (2017) conducted a review of AI-based regression models for electricity load prediction with a focus on only commercial buildings.

In response, Zhang et al., (2021) conducted a substantial review of review papers on energy consumption prediction models in academic literature and stipulated that most review studies focused on the comparison and juxtaposing of AI-based tools with many investigations and attention to the data area. Amongst the reviewed papers by Zhang et al., (2021), 110 concentrated on the tool aspect majorly and 50 considered the data aspect. Thus, the majority of review studies are conducted with a streamlined focus on the tools utilised in research studies. Notwithstanding, these review efforts are essential but there is still a deficit in review and technical studies that encapsulate building energy consumption prediction studies regarding the sizes of data (e.g., number of samples), type of energy (e.g., total electricity, natural gas), type of factors considered (e.g., meteorological, occupancy), type of building (e.g., commercial, residential) among others. A review of this sort is fundamental for subsequent well-informed application of tools in research as well as uncovering knowledge gaps and proffering future research focus toward advancing knowledge in the area of building energy consumption prediction.

3.4 Methodology

To select the appropriate tools for specific situations and suitable data conditions, a systematic analysis was conducted to develop a substructure for tool selection in the development of BEP models. Numerous tools have been utilized in the development of BEP models, such that it is basically impracticable to comparatively review all data-driven tools in one research. Therefore, several data-driven tools were selected for a systematic review based on popularity or promising potential. Specifically, a popular regression-based statistical tool: Linear Regression (LR) as noted in a comprehensive BEP model review by Mat Daut et al., (2017) and has been employed in BEP model deployment for the past two decades (Barakat et al., 1990; Divina et al., 2018a; Moghram and Rahman, 1989). Likewise, prevalent and promising AI-based tools as noted in a broad review by Zhang et al., (2021), Ahmad et al., (2014) and Zhao and Magoulès, (2012). and Amongst these is ANN which is stated as the most widely applied for decades(Krarti, 2003; Qiong Li et al., 2010; Shan et al., 2019), SVM which is showing promising results (Zhao and Magoulès, 2012), among others. This research conducts a systematic literature review of these tools.

A systematic literature review is a recognized approach of extensive literature to yield valid and reliable knowledge, considerably reducing chance and biases (Tranfield et al., 2003). A systematic literature review should adhere to a definite procedure to engender repeatability, clarity, and rigour. This research consisted of three phases: search for research publications (phase 1), implementation of inclusion and exclusion criteria (phase 2), implementation of bibliometric analysis (phase 3) and Systematic Analysis (phase 4). Subsequently, future research and knowledge gaps were derived from the result and presented. Figure 3.1 presents the framework of the primary step used in conducting this research.

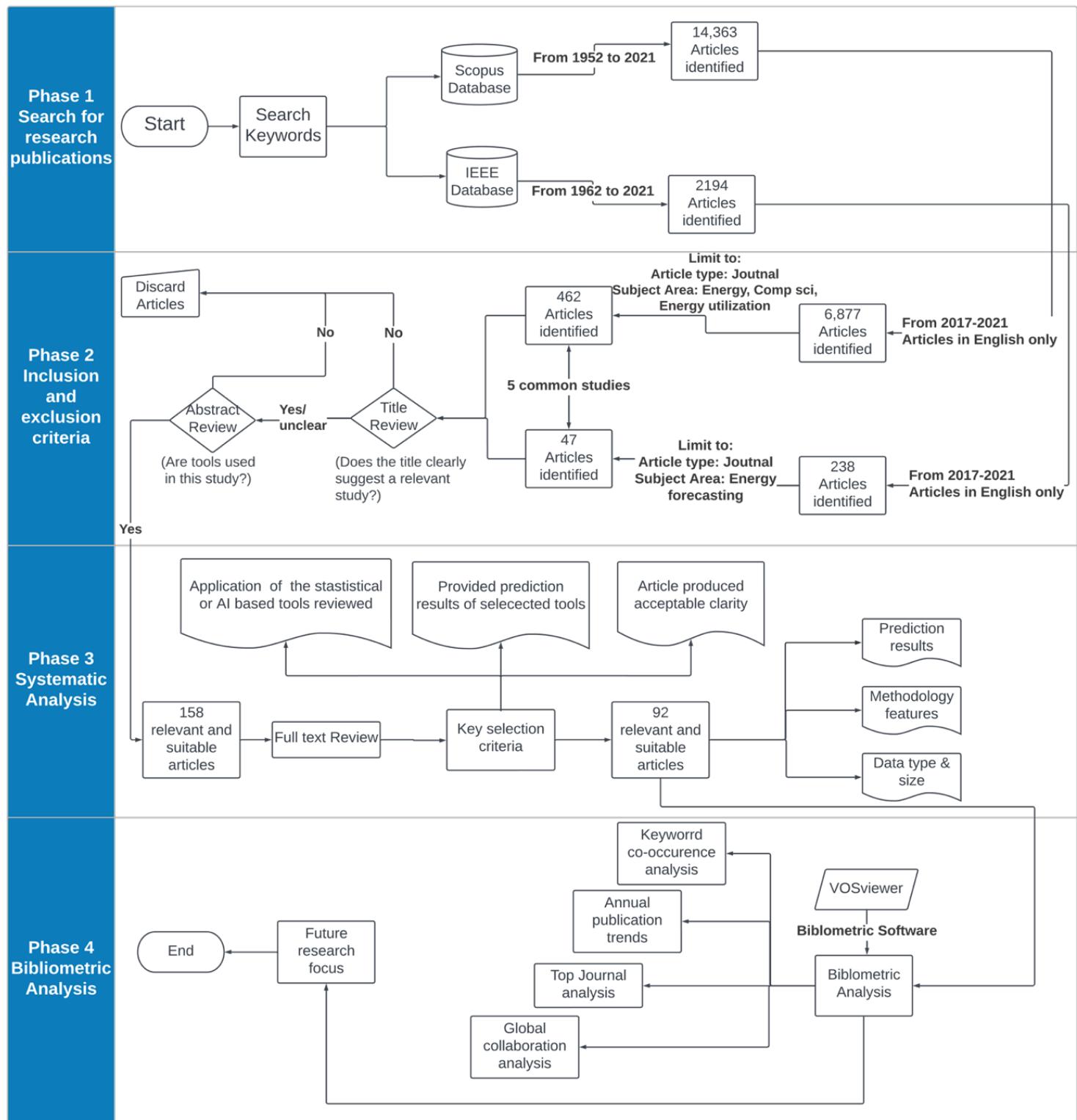


Figure 3.1: Framework of the primary steps of the methodology.

3.4.1 Data Collection

This section outlines the systematic approach employed to collect relevant studies for analysis. This process includes defining the research question and establishing clear inclusion and

exclusion criteria to ensure consistency and relevance. The process of searching for research articles and the inclusion and exclusion criteria are discussed below:

3.4.1.1 Search for Research Articles

As noted by Phillips and Barker, (2021), in a comprehensive research of review structure and form, the search phase and the strategy utilised for a systematic literature review is an imperative stage as it underpins the findings. This research adopted the search strategy of Amasyali and El-Gohary, (2018) which is one of the most recognised reviews in BEP with a citation of 813. Thus, as displayed in Figure 3.1 above, the first phase of this review includes the article search. The listed databases were considered: Engineering Village (EV), Scopus, Web of Science(WoS), Google Scholar, Institute of Electrical and Electronics Engineers (IEEE), and Information Service for Physics Electronics and Computing (INSPEC). EV, INSPEC and Scopus were considered because they are the most common databases amongst relatively high-impact factor energy journals such as Energy for Sustainable Development, and Energy among others. Although EV, INSPEC and WoS were not used due to accessibility limitations, the utilization of Scopus is considered sufficient for conducting a systematic literature review search owing to its high indexing rate and broad publication coverage (Debrah et al., 2022; Diirr and Santos, 2019). Also, Google Scholar was not used due to its production of boundless results with inconsistent accuracy and lack of effective filtering functionality as corroborated by (Falagas et al., 2008). To eliminate databases and geographic bias (Schlosser, 2007), two databases that encompassed studies from diverse countries around the world were used and this is confirmed and visualized in Figure 3.5.

The search keywords were carefully selected based on observation of the search outcome; various studies used synonyms ({“predict”, “forecast”}, {“power”, “load”, “consumption”}) as noun suffixes or verbs. These keywords were covered using Boolean operators (i.e., “OR”, and “AND”) to extract appropriate articles from the two databases (IEEE, Scopus) as displayed in Table 3.1 below.

Table 3.1: Database, keywords and articles search result

Databases	Query String	Results
Scopus	TITLE-ABS-KEY("building*" AND "energy" OR "electri*" OR "power" OR "load" AND "consumption" OR "performance" AND "forecast*" OR "Predict*") AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (SUBJAREA,"ENER") OR LIMIT-TO (SUBJAREA,"COMP")) AND (LIMIT-TO (DOCTYPE,"ar")) AND (LIMIT-TO (PUBYEAR,2021) OR LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR,2019) OR LIMIT-TO (PUBYEAR,2018) OR LIMIT-TO (PUBYEAR,2017)) AND (LIMIT-TO (LANGUAGE,"English")) AND (LIMIT-TO (EXACTKEYWORD,"Energy Utilization") OR LIMIT-TO (EXACTKEYWORD,"Forecasting"))	462
IEEE	("All Metadata":"forecast*" OR "All Metadata":"Predict*") AND ("All Metadata":"energy" OR "All Metadata":"electri*" OR "All Metadata":"power" OR "All Metadata":"load") AND ("All Metadata":"building") AND ("All Metadata":"consumption" OR "All Metadata":"performance") Filters Applied: Journals, 2017-2021, Energy forecasting (Subject Area)	47
	Manual selection as depicted in systematic analysis (phase 4)	92

3.4.1.2 Inclusion and Exclusion Criteria

The result shows that publications on energy consumption prediction increased significantly after the year 2017 which indicates that more research was conducted in this field from the year 2017 as shown in Figure 3.2, which is why 2017 was selected as the start year. Additionally, according to Zhang et al., (2021), the most recent review study was conducted in 2018 which reviewed the paper till the year 2017 (Seyedzadeh et al., 2018). Furthermore, the cut-off date for this search was December 2021. To enhance the validity of this research, solely journal

articles were employed, as they are deemed to be more credible. (Schlosser, 2007). Due to the often-inevitable interpretation cost constraints, 169 non-English articles were excluded from this research. The results emanated articles from other subject areas such as pharmacology (Reichert et al., 2019; Schlender et al., 2018), arts and humanities (Hung and Yang, 2018; Zheng et al., 2019) among others. Thus, the subject areas were further filtered to relevant subject areas namely energy utilization or forecasting and computer science. Subsequently, titles and abstracts were reviewed to detect unrelated articles. Several articles appeared in the search results that were tagged unrelated. This is because they did not focus on building energy consumption prediction but rather on topics such as occupancy forecasting (Li and Dong, 2017) and Environment performance evaluation (D'Amico et al., 2019) among others. The presence of unrelated articles in search results is subject to the use of certain keywords ("building", "prediction") in the abstract or title, hence they were discarded. Likewise, articles that did not employ any of the reviewed tools were removed (i.e., Cai et al., 2019; Fang et al., 2021). Consequently, bibliographic data for 92 articles were exported from both Scopus and IEEE before scanning or examining the full text, further consolidating the dataset for this research.

3.5 Bibliometric Analysis

Consequently, bibliometric analysis was conducted to better comprehend and examine knowledge areas. Accordingly, the selection of the most appropriate tool for various analysis is necessary (Darko et al., 2020). Many bibliometric or science mapping tools exist such as VOSviewer® (Van Eck and Waltman, 2020), Gephi® (Cherven, 2015), CiteSpace® (Chen, 2014) Sci2® (Weingart et al., 2010). However, of these tools, VOSviewer® is the most commonly utilized in research (Debrah et al., 2022; Olawumi et al., 2022) and it is considered to be user-friendly and one of the best tools for the bibliometric analysis (Boopathi and Gomathi, 2019; Saka and Chan, 2019). Hence VOSviewer® was applied in this research. VOSviewer® is a tool that offers the fundamental functions required for science mapping, visualization, and examination of bibliometric networks (Darko et al., 2020). The bibliometric analysis in this research was conducted for the following: annual publication trend analysis, keyword occurrence analysis, and Geographical publication or co-authorship analysis.

3.5.1 Publication Trends Analysis

Figure 3.2 shows the annual publication trend in the field of building energy consumption prediction from 2017 to 2021 which shows an exponential increase in this research field. Also,

Subject to the availability of data and computational competencies, research on the application of artificial intelligence or statistical tools for energy consumption prediction has received more attention since the year 2017 which expresses the reason for limiting the selection start year to 2017.

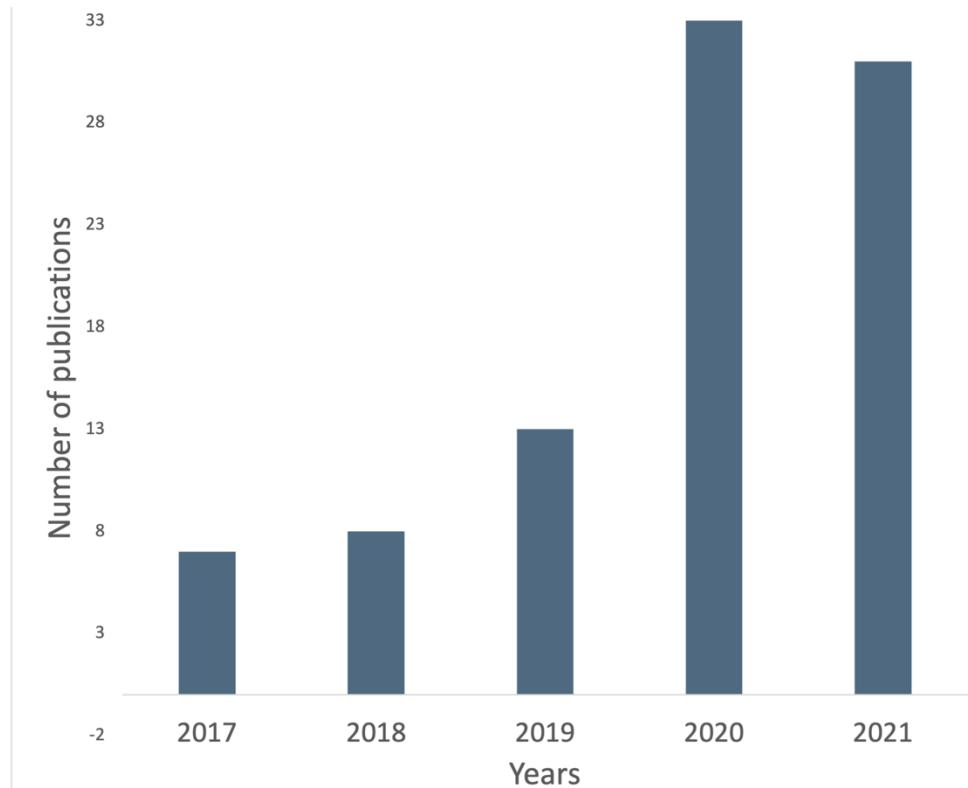


Figure 3.2: Annual publications of energy prediction articles

3.5.2 Keywords Co-Occurrence Analysis

The utilization of authors' keywords is commonly advocated for the identification of key research areas in the academic literature (Yin et al., 2019). VOSviewer was employed to visualize the keywords occurrence networks. The VOSviewer produces a science map based on distance, where the distance between two keywords represents the relational strength and the shorter distance indicates a more solid relationship (van Eck and Waltman, 2014). The size of the labels represents the rate of the keywords in relevant studies. VOSviewer visualizes them in various colours indicating different clusters. As noted in the publication trends (Figure 3.2), Statistical and AI tools applied in energy consumption prediction of buildings, have received more attention, resulting in 92 articles from Scopus and IEEE. The articles collectively generated a total of 951 keywords deduced using fractional counting. The number of occurrences was set to a minimum of 5, of which 82 keywords met the threshold. Subsequently,

a thesaurus file was created to consolidate similar keywords, for example, “buildings” was combined with “building”, “energy utilization” was combined with “energy use”, “build energy model” with “building energy modelling” and this was equally applied to other similar keywords. The generated network comprises of 4 clusters and 2078 links as displayed in Figure 3.3.

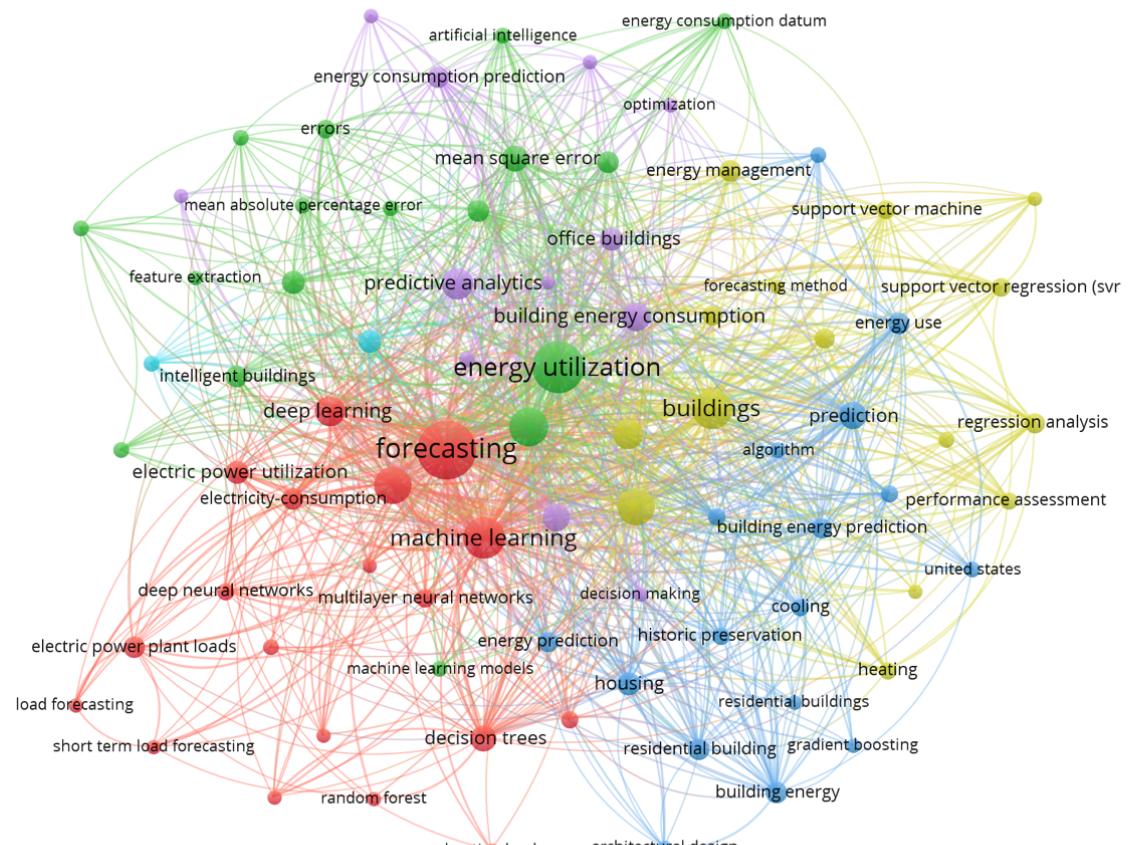


Figure 3.3: Keywords occurrence network

The top 20 keywords based on occurrences from the largest to the smallest were also displayed in Table 3.2. This table comprising of relevant information (average year) and Figure 3.3 visualizes how keywords or research areas are related and engendered a few findings as follows:

- 1) The top 20 most prominent keywords in the application of statistical and AI for building energy consumption prediction have an “average year published” of 2019 and 2020.
- 2) Specific keywords or research fields have received more attention than other fields. “Buildings”, “Energy utilization”, “Forecasting”, “Energy efficiency”, “Machine learning”, “Energy conservation” and “Learning systems” have been prolific in energy consumption prediction research. It is evident that “Machine Learning” has received

significant attention in the year 2020 (“Average year published”) often using the “Neural Network” tool.

Table 3.2: Top 20 keywords in energy consumption prediction

Keywords	Clusters	Number of links	Number of Occurrences	Average year published
Forecasting	1	81	83	2020
Energy utilization	2	81	64	2020
Machine learning	1	81	42	2020
Buildings	4	79	41	2019
Energy efficiency	2	78	36	2020
Energy conservation	4	77	35	2019
Learning systems	1	81	35	2020
Predictive analytics	5	71	25	2020
Deep learning	1	70	22	2020
Neural networks	4	79	22	2020
Building energy consumption	5	71	20	2020
Air conditioning	5	71	18	2020
Prediction	3	70	18	2020
Decision trees	1	61	17	2020
Mean square error	2	62	16	2020
Electric power utilization	1	61	14	2020
Housing	3	64	14	2020
Long short-term memory	2	54	14	2021
Office buildings	5	63	14	2019
Support vector machines	6	57	13	2020

3.5.3 Top Journal Analysis

It is important to analyse academic journals in specific research fields to avail authors with information on top outlets for publishing and readers with outlets to best search for resources. VOSviewer was utilized to generate the top journals in the field of building energy consumption prediction as presented and visualized in Table 3.3 and Figure 3.4 respectively. Table 3.3 shows the top 20 journals, the number of documents produced, the total number of citations for the journal documents and the strength of the links. Fractional counting was utilized, and it generated 136 journals. The minimum number of documents in each journal was set to 2, as performed in previous studies (Darko et al., 2020; Debrah et al., 2022). Thereafter, 45 journals met the threshold. The generated network comprises of 82 links as displayed in Figure 3.4. Table 3.3 shows that Applied Energy, Energy, Building and Environment, Energies, and IEEE. However, applied energy appears to account for 80 documents and 3709 citations.

Table 3.3: Top 20 journals and citations

Journals	Documents	Citations	Total link strength
Applied energy	80	3709	131
Energy	35	777	57
Building and environment	37	732	17
Energies	64	690	61
IEEE transactions on smart grid	4	420	4
Sustainable cities and society	17	407	17
Resources, conservation and recycling	2	323	5
IEEE access	18	216	11
Journal of cleaner production	9	185	5
IEEE internet of things journal	7	149	6
Solar energy	9	149	10
Building simulation	12	104	11
Neurocomputing	4	101	12
Energy conversion and management	8	93	5
Energy efficiency	8	86	4
Sensors (Switzerland)	3	79	0
Energy policy	2	70	1
Journal of building performance simulation	6	68	7
Science of the total environment	1	58	0
Renewable and sustainable energy reviews	2	43	12

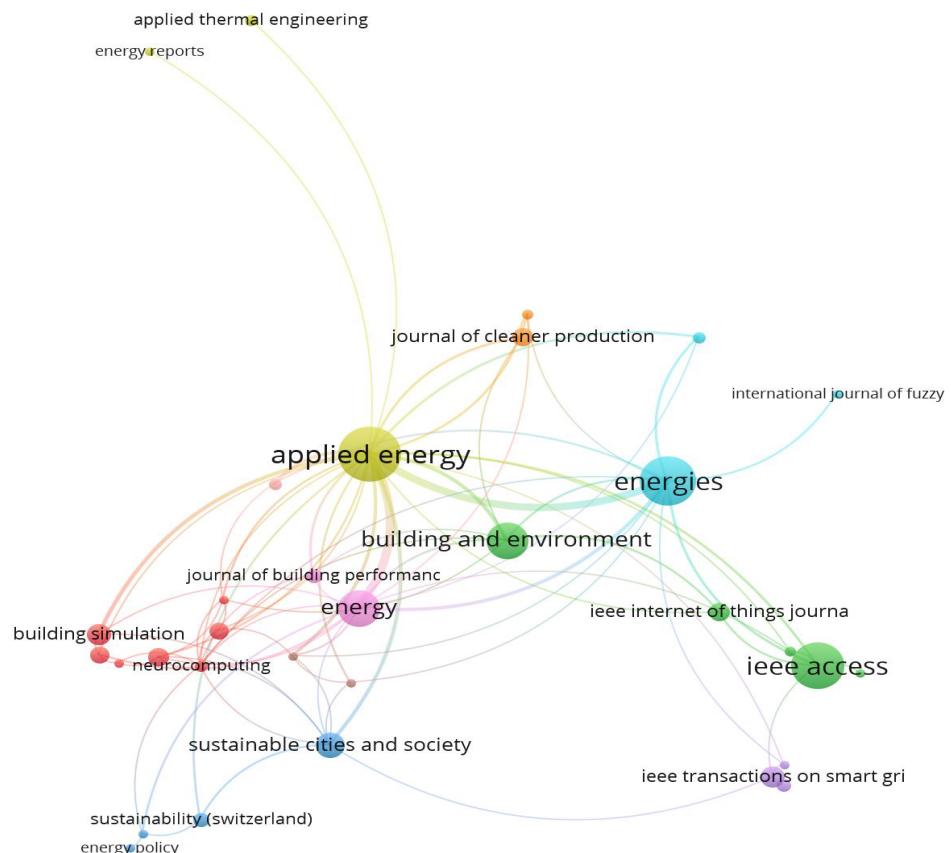


Figure 3.4: Journal citation network

3.5.4 Global Collaboration Analysis

In recent years, there lies a global increase in the research on the application of diverse methods for energy consumption prediction in buildings (Alduailij et al., 2021; Amasyali and El-Gohary, 2021). Though, Figure 3.5 displays the countries researching the application of statistical and AI-based tools for energy consumption prediction. Using Fractional counting, 41 countries were generated and the minimum number of documents of each country was capped at 2, resulting in 19 counties meeting the threshold.

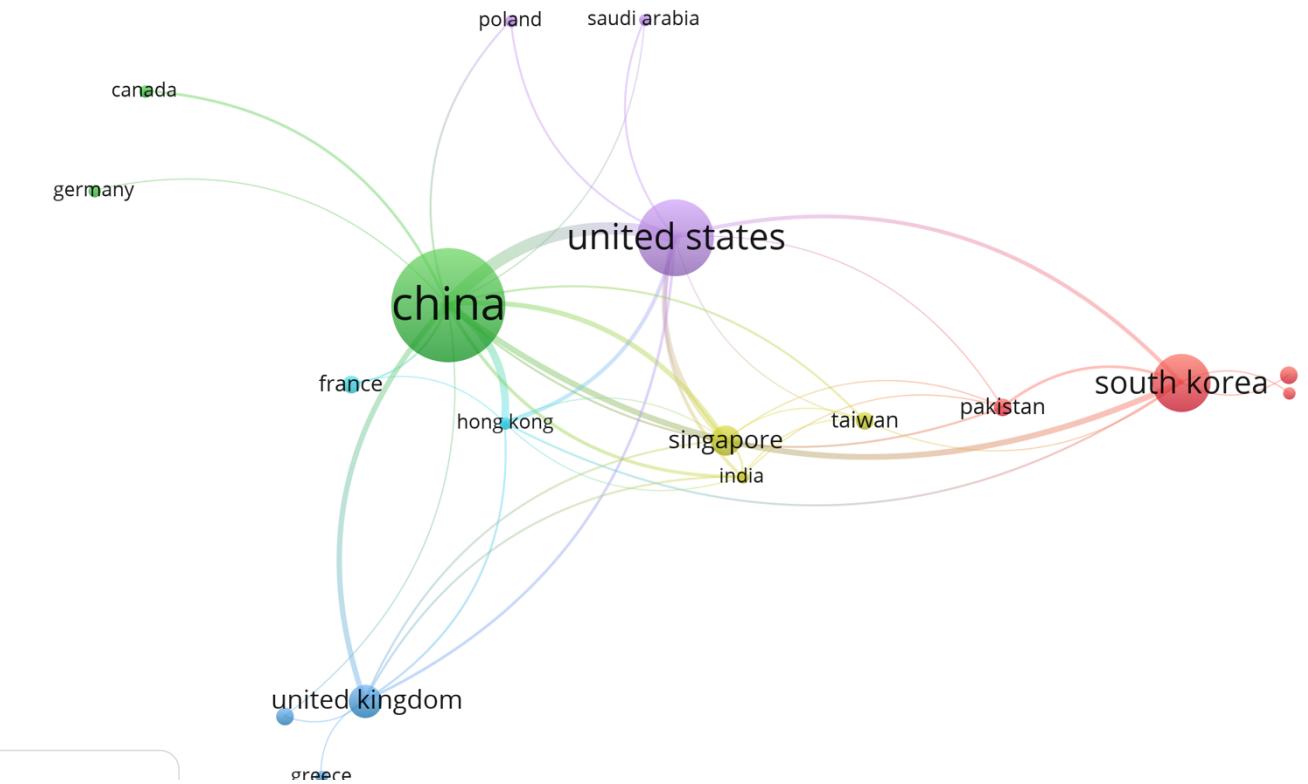


Figure 3.6 shows that the research on the application of AI-based methods for energy consumption prediction is more prominent in China and the United States (US). China has the highest number of publications and citations followed by the US. The high number of citations could represent the novelty and importance of the research. The pronounced research in these two countries is subject to the urgent need to understand the building energy efficiency (Khan et al., 2021; C. Li et al., 2017). For example, China has been the country with the world's largest population, leading to a significant increase in the energy consumed (Somu et al., 2020; R. Wang et al., 2018). The international energy outlook (2017) has projected that China will

account for one-fourth of the electricity consumed in buildings worldwide by the year 2040 (Capuano, 2019). Also, the US Energy Information Administration (EIA) projected that the US building sector will consume beyond 1.3% per year from 2018 to 2050 (EIA, 2020). Other countries such as the United Kingdom (UK), South Korea, and Singapore also made relatively significant contributions. Nonetheless, most countries have a growing number of publications and citations which indicates a good contribution to knowledge. However, countries such as Germany and Greece have several publications conceivably without novelty, leading to very low citations.

The positive relationship between novelty and citation is an established theory in academic research (Uzzi et al., 2013; Veugelers and Wang, 2019; Wang et al., 2017; X. Zhang et al., 2021). Tussen et al., 2000 conveyed that citations are key indicators of significant contribution to a specific field (Tussen et al., 2000). This is because novel studies often contribute a new perspective, insights and knowledge, which leads to high appreciation and citation by other researchers seeking to expand on existing knowledge. Bornmann et al., (2020) conducted an examination of 2000 research papers from a chemistry journal to investigate the relationship between citation count and the quality of manuscripts (Bornmann et al., 2012). It was concluded that citation count has a significant correlation with the quality of the research paper.

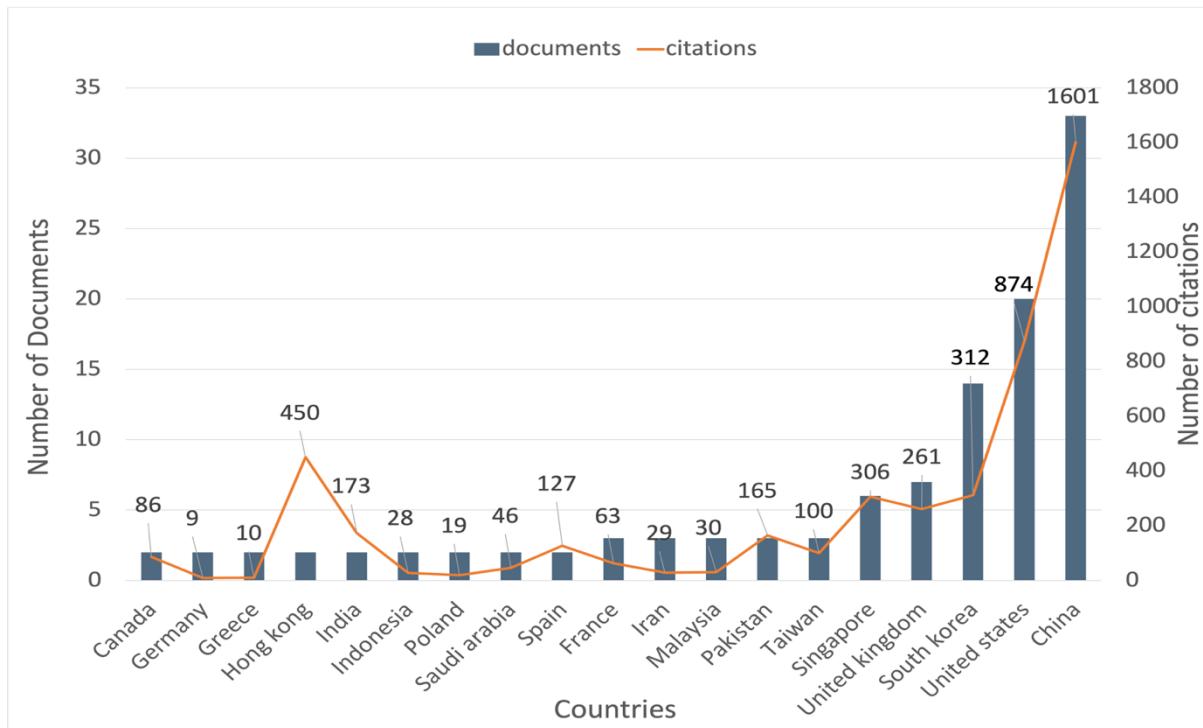


Figure 3.6: Global publications and citations distribution.

3.6 Systematic Analysis

To generate a comprehensive analysis of research for tool selection in the field of building energy consumption prediction, a systematic analysis of the chosen studies was conducted as displayed in phase 4 of Figure 1.1. After the articles were scanned and streamlined to 158 articles that were considered applicable, they were then assessed based on certain key selection areas. The following key selection criteria are as follows: articles that produced acceptable clarity, therefore, properly elucidate their methodology process, aim and objective, and conclusions. Articles that utilized at least one of the statistical or AI-based tools were selected for review. Subsequently, articles that clearly provided the prediction result of the tools employed. Thus, leaving the research with 92 relevant and suitable articles for review.

3.7 Statistical and AI-Based Tools

In this research, nine of the most utilised statistical and AI-based tools employed in the development of building energy prediction models were reviewed. These tools include Artificial Neural Networks (ANN), Extreme Gradient Boosting (XGB), Support Vector Machine(SVM), Random Forest (RF), Linear Regression (LR), Deep Neural Networks (DNN), Gradient Boosting Machines (GBM), Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM). A brief description of these tools is stated in Table 3.4 follows:

Table 3.4: Description of nine popular and promising statistical and AI tools

	Statistical/ AI tools	Description	References
1	Artificial Neural Network (ANN)	This is a powerful AI based tool for managing large and convoluted datasets (Sadeghi et al., 2020), which was first presented in 1949 (Gao et al., 2019). Essentially, ANN is devised to imitate how the neural system of a biological brain works (Ngarambe et al., 2020). However, certain criticisms have emerged based on its lack of model transparency (Ardjmand et al., 2016). ANN has the capacity for recognition and classification of patterns and learns from these historical patterns when analysing data (Ahmad et al., 2017b). Based on these merits, several researchers have employed ANN for developing a building energy use prediction model (Almalaq and Zhang, 2019; Hwang et al., 2020; Izidio et al., 2021; (Luo, 2023))	(Almalaq and Zhang, 2019; Chammas et al., 2019, p. 2; Divina et al., 2019; Feng and Zhang, 2020; Hwang et al., 2020; Sadeghi et al., 2020; Sha et al., 2019; Shan et al., 2019; Shapi et al., 2021)(Luo, 2023)

		Koukaras et al., 2021; Lee and Rhee, 2021; Shan et al., 2019 among others).	
2	Gradient Boosting Model (GBM)	GBM is an ensemble method centred on decision trees. GBM generates numerous models and delivers the optimum model that reduces the loss function. This is conducted by first developing various simple models, denoted as weak learners, and then merging them in a consecutive additive manner to achieve models with improved performance, denoted as strong learners (Ngarambe et al., 2020). GBM and RF have similar training processes, except for a primary difference, RF trains each decision tree independently, by utilizing random parameters, and merges the outcomes from all independent tree, while GBM trains decision trees one after the other, with the new trees bidding to reduce errors created by previous trees, a method recognised as boosting.	(Amber et al., 2018; Divina et al., 2018b; C. Fan et al., 2017; Hwang et al., 2020; Sadeghi et al., 2020)
3	Extreme Gradient Boosting (XGB)	XGB is one of the most utilised boosting tools in the machine learning field data in recent years (Kadkhodaei et al., 2020; Silvestro et al., 2017). XGB can be employed to provide solutions for both classification and regression problems. XGB is an advanced versions of gradient boosting machines (GBM). The relevant differences are its capacity to produce higher computation efficiency, and it employs a more regularized model formalization which tackles overfitting, which is recognised as one of the major problems of GBM (Divina et al., 2019). The iterative procedure of amending the prediction errors of previous models often enables XGB to elicit better performance than a single prediction algorithm (Cao et al., 2020). Due to its merits, XGB has attracted increased attention in the field of building energy prediction (Al-Rakhami et al., 2019; Divina et al., 2019; C. Fan et al., 2017; Khan et al., 2021).	(Chammas et al., 2019; Divina et al., 2018b; Kamel et al., 2020; Khan et al., 2021; Lee and Rhee, 2021)
4	Support Vector Machine (SVM)	SVM has received a considerable rate of recognition by researchers, due to the benefits of utilizing this method such as providing a solution to nonlinear problems while using a small training data size (Aversa et al., 2016; Mat Daut et al., 2017). This SVM is associated with the supervised learning techniques implemented for categorization and regression, introduced by Cortes and Vapnik, (1995). SVM was first applied in the development of an energy prediction model for monthly energy consumption in buildings by Dong et al., (2005) .	(C. Fan et al., 2017; Kontokosta and Tull, 2017; Li and Yao, 2020; Sha et al., 2019; Somu et al., 2020; G. Zhang et al., 2020)

5	Deep Neural Network (DNN)	DNN is similar to ANN as they are both designed to imitate the way the human brain functions (Ibraheem et al., 2017). They both have the capacity to learn highly convoluted patterns among variables (Ngarambe et al., 2020). However, the major distinction between ANN and DNN is the number of hidden layers in the model developed. An ANN model is habitually developed by utilizing a few hidden layers (usually three or less), while a model consisting of several hidden layers (Four or more) is recognised as a DNN (Olu-Ajayi et al., 2022a). A typical neural network model is often comprised of three layers (i.e., input layer, hidden layer, and output layer). Neurons are recognised as the primary units of the model, and they are connected through the layers.	(K. Amasyali and El-Gohary, 2021; Amber et al., 2018; K. Chen et al., 2019; C. Fan et al., 2017; Hwang et al., 2020; Sadeghi et al., 2020) (Ponta et al., 2022)
6	Random Forest (RF)	RF is a type of decision tree method based on ensemble techniques (Ngarambe et al., 2020). RF was created to prevail over certain weaknesses of standard decision tree methods such as lack of robustness (Ibrahim and Khatib, 2017). The robustness of RF is recognised as the highest form adaptation to various training samples with a minimum standard deviation of the error (Wang et al., 2019). RF is considered to produce fast training time and is dominant for resolving high-dimensional data and convoluted problems (Zhu et al., 2020). Particularly, RF, have been employed in the research to predict energy consumption in buildings (Chammas et al., 2019; Divina et al., 2018a; C. Fan et al., 2017; Khan et al., 2021).	(Chammas et al., 2019; Divina et al., 2018b; C. Fan et al., 2017; Khan et al., 2021; Kontokosta and Tull, 2017; Pan and Zhang, 2020)
7	Linear Regression (LR)	LR is a statistical method often utilised for time series prediction. The fundamental dynamics behind linear regression is to attempt to detect the relationship between two or more variables. LR employs a basic linear equation ($Y = a + bX$) to signify the relationship between the independent and dependent variables as denoted as (X) and (Y) respectively. In the event that several independent variables are utilised to calculate the value of a dependent variable, the technique is known as multiple linear regression. Therefore, the goal in such a situation is to model the linear association between multiple independent variables and a dependent variable by means of an equation such as $Y = a + b_1X_1 + \dots + b_nX_n + \epsilon \quad (1)$ <p>Where: ϵ is the difference between the observed and predicted value $X_{i, 1 \leq i \leq n}$ are the n explanatory variables.</p>	(Chou and Tran, 2018; Divina et al., 2018a; Hribar et al., 2019; Kamel et al., 2020; Kontokosta and Tull, 2017; Li and Yao, 2020)

8	Auto-Regressive Integrated Moving Average (ARIMA):	ARIMA is an advanced version of a basic AutoRegressive Moving Average (ARIMA) with inclusion of concept of integration. This method can be used for time series prediction, Hence, it has been employed for the development of building energy consumption prediction model (Hwang et al., 2020; D.M.F. Izidio et al., 2021; Shan et al., 2019; Somu et al., 2020). Despite their acceptance and recognition, these models are unable to apprehend complex interactions in non-linear data. Machine Learning (ML) approaches have been noted as the more appropriate approach to managing such complexity.	(Almalaq and Zhang, 2019; Chou and Tran, 2018; Divina et al., 2018b; Hwang et al., 2020; Somu et al., 2021, 2020) (Pora et al., 2022)
9	Long Short-Term Memory (LSTM):	LSTM is a deep learning tool that has become an attractive method for energy consumption prediction (Khan et al., 2021; Y. Liu et al., 2020a; Somu et al., 2021; J. Wang et al., 2021). The application of LSTM models may not yield satisfactory prediction results, not only the presence of noisy data but also because of the naive choice of its hyperparameter values (Bouktif et al., 2020).	(Chou and Tran, 2018; Hwang et al., 2020; Khan et al., 2021; Kong et al., 2019; C. Liu et al., 2020; Shan et al., 2019; Somu et al., 2021)

3.7.1 Relevant Criteria

In consideration of the development of an effective energy use prediction model, there are several criteria to satisfy. However, the required criteria differ, dependent on the condition and goal of the model developer. For example, accuracy is of great importance when developing a BEP model however the accuracy rate is highly correlated to the input/output, and data size amongst other criteria(Fathi et al., 2020b; Goyal et al., 2020; Runge and Zmeureanu, 2019). Also, there is no combination of criteria that is considered the best for accomplishing the most accurate AI-based tool prediction, as there is no general scope that delivers such a global comparison (Fathi et al., 2020b). A comprehensive review of key articles in the field of building energy consumption prediction engendered six (6) criteria based on popularity and importance.

Table 3.5 describes the six different criteria:

Table 3.5: Description of six key criteria in the field of energy prediction.

Criteria	Description	Reference
1 Error rate	This refers to the error rate between the predicted energy value and actual energy values.	(Amasyali and El-Gohary, 2018)
2 Building type	This relates to the tool's performance in the prediction of different types of building such as residential and non-residential.	(Molina-Solana et al., 2017; Wang and Srinivasan, 2017)

3	Energy type	This relates to the tool's performance in the prediction of different energy types consumed in buildings. These energy types include Heating, Cooling, Natural gas, among others.	(Wang and Srinivasan, 2017)
4	Temporal granularities	The related performance of the tool on different time scale categories which includes Annual, Monthly, Weekly, Daily, Hourly, and Sub-hourly.	(Amasyali and El-Gohary, 2018; Lazos et al., 2014)
5	Data size	This refers to the size of data required or appropriate for a tool to produce optimal performance.	(Amasyali and El-Gohary, 2018; Li and Wen, 2014)
6	Variable selection	This refers to the variable or feature selection methods needed for the tool to produce optimal performance.	(Raza and Khosravi, 2015)

3.8 Result and Discussion

This systemic literature review gives the result and findings of the systematic literature review based on the acknowledged criteria and visualizes these results using tables and statistical diagrams. The scope of this review was categorized based on five types of energy (Heating, Cooling, Natural gas, Total electricity, Total building energy), three types of building (Residential, Commercial, Educational & Research), 2 types of input data (Meteorological and Occupancy data) and temporal granularities (Annual, Monthly, Weekly, Daily, Hourly, Sub-hourly). The comparison of the tool's performance in relation to the reviewed studies will be conducted using tables and diagrams. The result was used in examining each tool based on the specific criteria. The identified criteria were evaluated and presented based on the building energy prediction tools and studies. For instance, to evaluate the error rate, the type of error evaluation method employed and what it represents (i.e., For Mean Absolute Error (MAE) - the closer the outcome is to zero, the more enhanced the performance and the higher the outcome, the worse the performance).

The reviewed studies explored the development of models for energy consumption prediction for residential, commercial, or educational/research buildings based on various temporal granularities for various energy types. The proportion of reviewed studies in relation to building types, temporal granularity, energy types and tools utilized are visualised using pie charts in Figure 3.7 below.

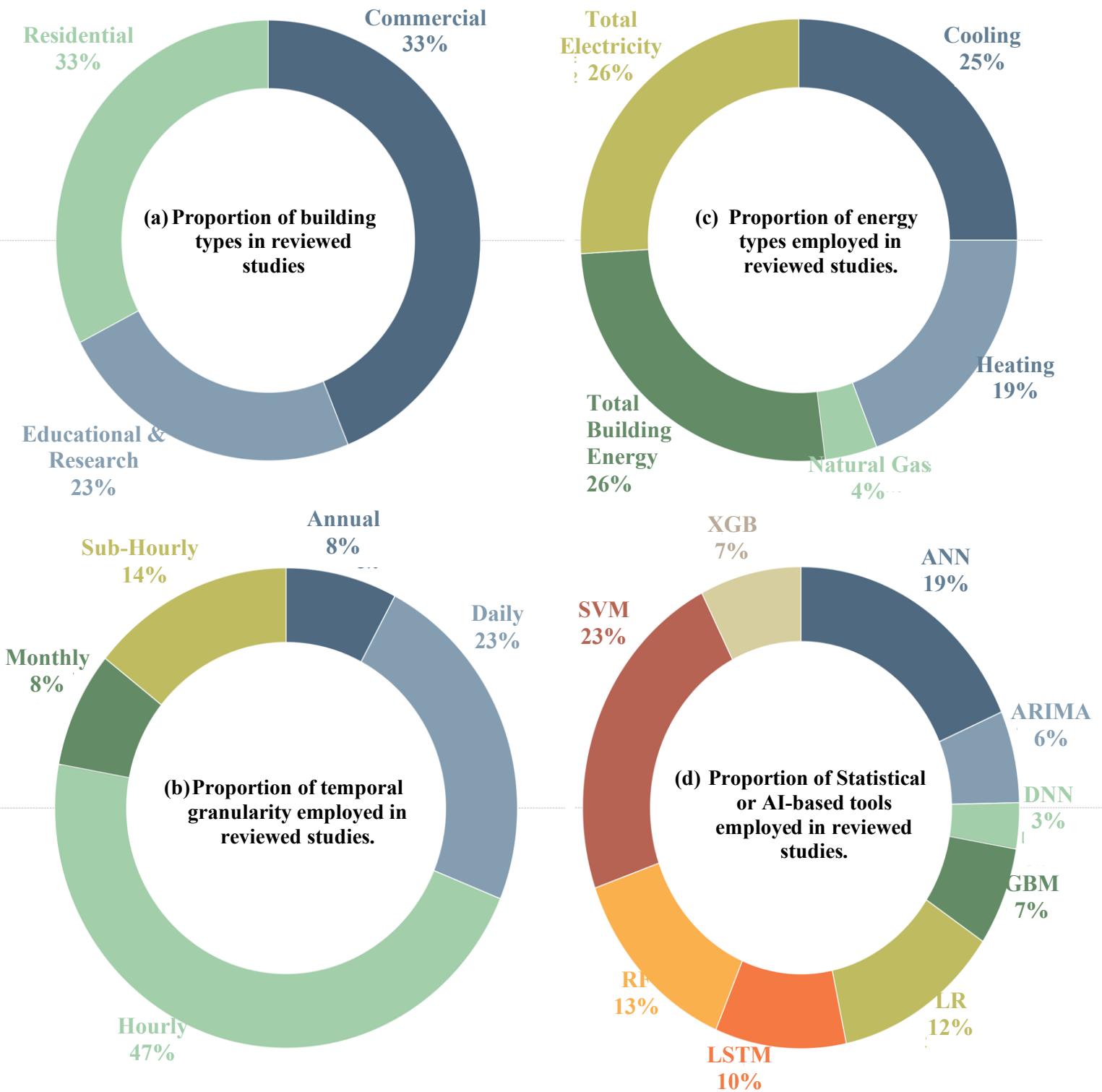


Figure 3.7: The proportion of reviewed studies based on (a) building types, (b) temporal granularity, (c) energy types, (d) Statistical or AI-based tools

Figure 3.7a – 3.7d shows that only 33% of reviewed studies concentrated on residential buildings while other studies concentrated on non-residential buildings (44% on commercial buildings, and 23% on educational and research buildings). Regarding temporal granularity, a significant portion of studies concentrated on hourly energy consumption prediction which totals 47% while 8%, 8%, 23% and 14% of studies concentrated on the yearly, monthly, daily and sub-hourly consumption respectively. The various energy types were explored in the reviewed studies with 26%, 26%, 25%, 19%, and 4% of total building energy, total electricity, cooling, heating, natural gas, respectively. Generally, 76% of studies employed meteorological data as the features in the development of energy consumption prediction model while 24% utilized occupancy data. Various statistical and AI based models have been implemented in the reviewed research, however, the most utilized is SVM with 23%, following this is 19%, 13%, 12%, and 10% of ANN, RF, LR and LSTM respectively while the least utilized are XGB, GBM, ARIMA and DNN with 7%, 7%, 6% and 3% respectively. The scope of all reviewed studies is shown in Table 3.6, based on types of building, type of energy, temporal granularity, type of energy consumption, input data, purpose of research, data size and performance.

Table 3.6: Data properties, purpose, and performance of statistical and AI based tools employed in reviewed studies.

S/N	Reference	Purpose of study	Tool type	Feature selection	Building type	Temporal granularity	Energy type	Input features	Data size	Performance measures
1	Divina et al., (2019)	Model comparison	ANN							0.99 (MAE)
			RF							0.97 (MAE)
			LR		Educational & Research	Daily	Total Electricity	Meteorological conditions	13 instances	0.98 (MAE)
			GBM							0.97 (MAE)
			ARIMA							1.3 (MAE)
			XGB							0.97 (MAE)
2	Hribar et al., 2019	Model comparison	LR	Spearman's rank	Residential	Hourly	Natural Gas	Meteorological conditions		1.1 (MAE)
3	Kontokosta and Tull, (2017)	Model comparison	SVM		Residential				20000 instances	0.75 (MAE)
			RF	Stepwise selection						0.79 (MAE)
			LR		Commercial	Annual	Natural Gas	Occupancy		0.8 (MAE)
4	Somu et al., (2021)	Propose Hybrid Model (kCNN-LSTM)	ANN ARIMA LSTM		Educational & Research	Daily	Total Building Energy		4 instances	0.18 (MAE) 0.1 (MAE) 0.34 (MAE)
5	Somu et al., (2020)	Propose eDemand model	SVM ARIMA		Educational & Research		Total Building Energy	Meteorological conditions	1 instance	0.1 (MAE) 0.34 (MAE)
6	Liu et al., (2020)	Propose hybrid model based on the Holt-Winters (HW) and Extreme Learning Machine (ELM) network	LSTM		Residential	Sub-Hourly	Total Electricity		1 instance	0.086 (MAE)
7	Li and Yao, (2020)	Compare Data-Driven models	ANN							0.34 (MAE)
			SVM		Residential	Annual	Heating Cooling	Meteorological conditions	200 instances	0.71 (MAE)
			LR					Occupancy		1.36 (MAE)

		SVM								24.467 (CVRMSE)
		RF	engineering, statistical, and structural feature extraction methods							34.95 (CVRMSE)
		LR		Educational & Research	Daily	Cooling	Meteorologic al conditions	1 instance		43.9 (CVRMSE)
		DNN								24.567 (CVRMSE)
		GBM								23.95 (CVRMSE)
		XGB								23.67 (CVRMSE)
8	Fan et al., (2017)	Model comparison								
9	Sha et al., (2019)	Proposed a simplified data-driven model for predicting the energy consumption of an HVAC system in a building	ANN							27.1 (CVRMSE)
10	Zhang et al., (2020)	Propose a new method	SVM		Sub-Hourly			2 instances		38.7 (MAE)
11	Al-Rakhami et al., (2019)	proposed an ensemble learning approach	XGB	Residential		Heating				0.23 (MAE)
12	Chammas et al., (2019)	Propose a system based on Multilayer Perceptron (MLP)	ANN SVM RF LR GBM	Commercial	Total Building Energy	Meteorologic al conditions	2 instances			29.55 (MAE) 31.36 (MAE) 31.85 (MAE) 51.97 (MAE) 35.22 (MAE)
13	Khan et al., (2021)	Proposed a spatial and temporal ensemble forecasting model	RF GBM XGB LSTM		Hourly	Total Electricit y	4 instances			0.065 (R2) 0.046 (R2) 0.066 (R2) 0.901 (R2)
14	Sadeghi et al., (2020)	Applied deep neural networks (DNNs) to forecast HLs and CLs	ANN DNN	Residential	Heating		768 instances			2.42 (MAE) 1.22 (MAE)
15	Feng and Zhang, (2020)	Model comparison	ANN SVM RF GBM	Educational & Research	Sub-Hourly	Total Building Energy	Meteorologic al conditions	13 instances		1.46 (MAE) 1.57 (MAE) 1.05 (MAE) 1.34 (MAE)
16	Wang et al., (2021)	Proposed a deep convolutional neural network based on ResNet	LSTM	Commercial	Hourly	Total Building Energy	Meteorologic al conditions	3 instances		3.18 (MAE)
17	Li et al., (2017)	Model comparison	SVM LR	Commercial	Sub-Hourly	Total Building Energy		1 instance		16.36 (MAE) 25.75 (MAE)
18	Pan and Zhang, (2020)	Proposed categorical boosting (CatBoost)- based predictive method	RF	Commercial	Annual	Total Building Energy				7.5 (RMSE)
		LR	Multiple (Pearson correlatio n LR, Backward Eliminatio n Correlatio n Analysis)							0.057 (RMSE)
19	Kamel et al., (2020)	Explores the effect of feature selection	XGB	Residential	Heating Cooling	Meteorologic al conditions	1 instance			0.038 (RMSE)
20	Li et al., (2018)	Proposed a modified deep belief network (DBN) based hybrid model	SVM	Commercial	Sub-Hourly		2 instances			29.16 (MAE)

21	Ahmad et al., (2018)	Investigates the accuracy and generalisation capabilities of deep highway networks (DHN) and extremely randomized trees (ET)	SVM	RF & Extra Tree	Commercial	Hourly	Heating Cooling	Meteorologic al conditions Occupancy	1 instanc e	3.09 (MAE)
22	Jang et al., (2019)	Extracting major variables & machine learning model optimization	ANN	Random forest	Educational & Research	Hourly	Heating	Meteorologic al conditions	6 instances	9.52 (MAE)
23	Shao et al., (2020)	proposed a novel domain fusion deep model based on CNN, LSTM, and discrete wavelet transform (DWT)	DNN		Commercial	Hourly	Total Electricit y		2 instances	1.04 (MAPE)
24	Cao et al., (2020)	Model comparison	SVM RF LR XGB		Commercial	Daily	Total Electricit y	Meteorologic al conditions Occupancy	1 instanc e	13.68 (CVRMSE) 12.57 (CVRMSE) 17.65 (CVRMSE) 12.81 (CVRMSE)
25	Liu et al., (2020)	Used SVM method to predict energy consumption	SVM	Predictive sample	Commercial	Daily	Total Building Energy	Meteorologic al conditions	1 instanc e	0.046 (RMSE)
26	Shapi et al., (2021)	Model comparison	ANN SVM		Commercial		Total Electricit y		1 instance	8.65 (MAE) 6.24 (MAE)
27	Pinto et al., (2021)	Model comparison	SVM RF GBM		Commercial	Hourly	Total Electricit y	Meteorologic al conditions	1 instanc e	5.82 (MAE) 6.11 (MAE) 6.07 (MAE)
28	Almalaq and Zhang, (2019)	Proposed a novel hybrid prediction approach based on the evolutionary deep learning (DL) method	ANN		Residential		Total Electricit y		2 instances	0.186 (MAE)
29	zidio et al., (2021)	Proposed an Evolutionary Hybrid System (EvoHyS)	ARIMA LSTM ANN SVM LR ARIMA LSTM		Commercial Residential	Hourly	Total Building Energy		1 instanc e	0.196 (MAE) 0.182 (MAE) 0.057 (MAE) 0.07 (MAE) 0.062 (MAE) 0.091 (MAE) 0.029 (MAE)
30	Shan et al., (2019)	Proposed a gravity gated recurrent unit electricity consumption model	ANN SVM ARIMA LSTM		Commercial	Hourly	Total Electricit y		1 instanc e	68.31 (MAPE) 11.68 (MAPE) 4.17 (MAPE) 3.72 (MAPE)
31	Hwang et al., (2020)	Proposed a two-step approach to develop a highly accurate electricity consumption prediction model	ANN SVM DNN ARIMA LSTM		Commercial	Daily Monthly	Total Electricit y	Meteorologic al conditions	28 instances	26.88 (MAPE) 23.7 (MAPE) 14.65 (MAPE) 26 (MAPE) 14.52 (MAPE) 0.057 (MAE) 0.07 (MAE) 0.062 (MAE) 0.091 (MAE)
32	Chou and Tran., (2018)	Evaluate the effectiveness of statistical and ML models.	ANN SVM LR ARIMA		Residential	Daily	Total Building Energy	Meteorologic al conditions	1 instanc e	

33	Syed et al., (2021)	Proposed a novel hybrid deep learning model	LR LSTM	Residential	Total Building Energy	Meteorologic al conditions	1 instanc e	74.26 (MAE) 25.23 (MAE)	
34	Lee and Rhee, (2021)	Adopted transfer learning and meta learning integrated into DNN	ANN XGB LSTM	Residential	Hourly	Total Electricit y		0.27 (RMSE) 0.272 (RMSE) 0.2857 (RMSE)	
35	Amber et al., (2018)	Model comparison	ANN SVM DNN	Educational & Research	Daily	Total Electricity	Meteorologic al conditions	17 (MAE) 24.11 (MAE) 27 (MAE)	
36	Koukaras et al., (2021)	Proposed a novel approach for One-Step-Ahead Energy Load Forecasting (OSA-ELF)	ANN SVM LR GBM	Residential	Hourly	Total Electricity	1 instanc e	0.668 (MAE) 0.796 (MAE) 0.68 (MAE) 0.6987 (MAE)	
37	Robinson et al., (2017)	Proposed a novel method	ANN SVM RF LR XGB	Commercial	Annual	Total Building Energy Natural Gas		0.28 (MAE) 0.25 (MAE) 0.29 (MAE) 0.29 (MAE) 0.24 (MAE)	
38	Amasyali and El-Gohary, (2021)	Proposed a machine-learning method for predicting energy consumption in consideration occupant-behavior	ANN DNN	Commercial	Hourly	Total Building Energy Cooling	Meteorologic al conditions Occupancy	1152 instances 24.075 (MAE)	
39	Guo et al., (2018)	Model comparison	SVM LR	Correlatio n analysis & LASSO	Commercial	Sub-Hourly	Heating	Meteorologic al conditions	1 instanc e 3.1264 (MAE) 2.5586 (MAE)
40	Pham et al., (2020)	Proposed a RF model	RF		Hourly	Total Building Energy		5 instances 2.77667 (MAE)	
41	Borowski and Zwolińska, (2020)	Compared neural networks and support vector machines.	ANN SVM	Commercial	Sub-Hourly	Cooling	Meteorologic al conditions	1 instanc e 11.262 (MAE) Occupancy 13.715 (MAE)	
42	Kim et al., (2020)	Model comparison (Linear regression) and artificial neural network (ANN) algorithms.	ANN	Educational & Research	Hourly		Meteorologic al conditions	1 instanc e 7.04 (CVRMSE)	
43	Eseye and Lehtonen, (2020)	Proposed a novel model based on the integration of empirical mode decomposition (EMD), imperialistic competitive algorithm (ICA), and SVM.	LR ANN SVM	Residential Commercial Educational & Research	Hourly	Heating	Meteorologic al conditions	4 instances 4.18 (CVRMSE) 64.4 (MAPE) Occupancy 19.87 (MAPE)	
44	Ullah et al., (2020)	Proposed an intelligent hybrid methods that integrates CNN with a Multi-layer Bi-directional Long-short Term Memory (M-BDLSTM) method	LSTM	Residential		Total Building Energy		1 instanc e 0.4008 (MAE)	
45	Feng et al., (2021)	Models comparison	SVM XGB	Correlatio n analysis	Residential	Hourly	Cooling	Meteorologic al conditions	391 instances 0.643 (MAE) 0.294 (MAE)

46	Mohammed et al., (2021)	Models comparison	RF XGB	Correlation analysis	Residential	Heating Cooling	768 instances	0.7625 (MAE) 0.7225 (MAE)
47	Seyedzadeh et al., (2019)	Investigated the accuracy of most popular ML models	ANN SVM RF XGB	Statistics based method, PCA, fully connected autoencoders (AEs), convolutional autoencoders (CAEs) and generative adversarial networks (GANs)	Residential	Heating Cooling	Meteorological conditions 2 instances	2.8155 (MAE) 2.70575 (MAE) 2.8025 (MAE) 1.939 (MAE)
48	Fan et al., (2019)	Investigates the performance of various deep learning methods	XGB	Hourly	Cooling	Meteorological conditions	1 instance	23.22 (CVRMSE) 21.82 (CVRMSE) 30.78 (CVRMSE) 18.72 (CVRMSE)
49	Pino-Mejías et al., (2017)	Model comparison	ANN LR	Commercial	Heating Cooling			0.99 (MAE) 0.948335 (MAE)
50	Fu, (2018)	Proposed a deep learning based hybrid approach	SVM	Commercial	Sub-Hourly	Cooling	Meteorological conditions	22.665 (MAE)
51	Zhong et al., (2019)	Proposed a novel vector field-based support vector regression method	SVM LR	Commercial	Hourly	Total Building Energy	Meteorological conditions 1 instance	38.40055 (MAE) 31.6698 (MAE)
52	Shao et al., (2020)	Model comparison	GBM SVM	Commercial	Hourly	Cooling	Meteorological conditions 1 instance	31.7424 (MAE) 4.83 (MSE)
53	Moon et al., (2021)	Proposed a hybrid short-term load forecast model	ANN SVM RF GBM	Educational & Research	Daily	Total Electricity	Meteorological conditions 68 instances	3.990667 (MAPE) 4.3223 (MAPE) 4.34733 (MAPE) 5.215667 (MAPE)
54	Wang et al., (2020)	proposed a novel improved integration model (stacking model)	SVM RF XGB	Educational & Research	Hourly	Heating	Meteorological conditions 2 instances	18.095 (MAE) 20.63 (MAE) 16.05 (MAE)
55	Bouktif et al., (2020)	Proposed an optimal configuration of an LSTM model	ANN SVM RF	Commercial	Sub-Hourly	Total Electricity		1.311 (CVRMSE) 0.835 (CVRMSE) 0.805 (CVRMSE)
56	Ullah et al., (2020)	Proposed a Hidden Markov Model	ANN SVM	Residential	Hourly	Total Electricity	Occupancy	2.7 (MAE) 2.79 (MAE)
57	Zhou et al., (2020)	Model comparison	ARIMA LSTM	Educational & Research	Daily	Cooling		10.37 (MAPE) 8.68 (MAPE)

58	Wang et al., (2019)	Model comparison	SVM RF XGB	Residential	Hourly	Heating	Meteorologic al conditions Occupancy	0.25 (MAE) 0.2 (MAE) 0.21 (MAE)
59	Kim et al., (2020)	Model comparison	ANN LR	Educational & Research	Hourly	Total Electricit y	Meteorologic al conditions	1 instance 6.275 (CVRMSE) 8.29 (CVRMSE)
60	Safa et al., (2017)	Model comparison	ANN LR	Educational & Research	Monthly	Total Building Energy	Meteorologic al conditions	4 instances 0.8325 (R2) 0.7775 (R2)
61	Irfan et al., (2021)	Model comparison	SVM RF	Residential	Hourly	Heating Cooling	768 instances 2.145 (MAE) 0.965 (MAE)	
62	Hosseini and Fard, (2021)	Model comparison	RF Univariate regression algorithm	Residential		Heating Cooling	22 instances 1.128 (MAE)	
63	Groß et al., (2021)	Model comparison	ANN SVM RF LR ARIMA LSTM	Sequential forward selection	Residential Commercial Educational & Research	Daily	Total Electricit y	3 instances 0.371667 (MAE) 0.3483 (MAE) 0.41167 (MAE) 0.42 (MAE) 0.385 (MAE) 0.4016667 (MAE)
64	Liu et al., (2021)	Proposed an approach to predict building energy consumption	SVM RF	Correlatio n analysis	Educational & Research	Total Building Energy Occupancy	Meteorologic al conditions 1 instance 0.426 (RMSE) 0.015 (RMSE)	
65	Goudarzi et al., (2021)	Proposed an enhanced hybrid model based on ARIMA and Imperialist Competitive Algorithm (ICA).	ARIMA		Educational & Research	Total Building Energy	1 instance 1.5766 (MAE)	
66	Ding et al., (2020)	Model comparison	RF GBM	Correlatio n analysis, K-means clustering, and discrete wavelet transform (DWT)	Commercial	Hourly	Cooling	Meteorologic al conditions 1 instance 37.975 (R2) 40.725 (R2)
67	Liu et al., (2021)	Model comparison	ANN	Residential	Sub-Hourly	Heating	Meteorologic al conditions	1 instance 4.15 (RMSE)
68	Hadri et al., (2021)	Explored three main approaches (univariate, multivariate and multistep)	XGB LSTM	Commercial	Daily	Total Electricit y Occupancy		7.988 (MAE) 10.304 (MAE)
69	Gao et al., (2021)	leveraged deep learning models to predict energy consumption	ANN SVM RF XGB	Commercial	Hourly	Total Building Energy	16 instances 0.641 (MAE)	0.566 (MAE) 0.817 (MAE) 0.823 (MAE)

70	Gao et al., (2020)	Proposed a novel ensemble prediction model	ANN SVM GBM	Residential	Daily	Heating	3 instances	50.773 (MAE) 64.178 (MAE) 59.99667 (MAE)
71	Xie et al., (2020)	Proposed a hybrid hour-ahead forecast model	SVM LSTM ARIMA	Commercial	Hourly	Cooling		0.316 (MAE) 0.354 (MAE) 21.73 (MAE)
72	Lin et al., (2021)	predict the air conditioning power consumption and lighting power consumption, respectively	LSTM	Commercial	Daily	Total Electricity	Meteorologic al conditions	1 instance 15.34 (MAE)
73	Alduailij et al., (2020)	Proposed statistical, time series, and machine learning techniques	ANN LR ARIMA LSTM	Educational & Research	Daily	Total Electricity	Meteorologic al conditions	5 instances 1.6892 (MAE) 26.0246 (MAE) 1.0855 (MAE) 1.3804 (MAE)
74	Liao et al., (2020)	Model comparison	ANN SVM RF LR DNN LSTM	Commercial	Monthly	Total Electricity	Meteorologic al conditions	1 instance 0.57 (MAE) 0.77 (MAE) 0.88 (MAE) 0.57 (MAE) 0.7 (MAE) 0.73 (MAE)
75	Kim et al., (2020)	Model comparison	ANN	Commercial	Hourly	Cooling	Meteorologic al conditions	1 instance 19.486 (MAE)
76	Guo et al., (2020)	Used 40 weather factors in energy prediction	SVM RF GBM stepwise, least angle regression (Lars), and Boruta algorithms	Commercial		Total Building Energy	Meteorologic al conditions	1 instance 0.07694767 (MAE) 0.09786 (MAE) 0.0949 (MAE)
77	Zhang et al., (2021)	proposed a clustering decision tree algorithm to identify the building operation conditions.	ANN SVM RF GBM	Commercial			Meteorologic al conditions	1 instance 0.682 (R2) 0.648 (R2) 0.816 (R2) 0.728 (R2)
78	Nie et al., (2021)	Proposed an innovative energy consumption prediction model	SVM ARIMA	Residential		Total Electricity		89.49 (MAE) 87.4 (MAE)
79	Elbeltagi and Wefki, (2021)	Proposed a methodology based on ANNs to enhance the prediction of energy	ANN	Residential	Annual	Total Building Energy	Meteorologic al conditions	1 instance 0.96 (MAE)
80	Chen et al., (2020)	Proposed new meta ensemble learning method	SVM	Commercial	Monthly	Cooling	Meteorologic al conditions	1 instance 4.487 (MAE)
81	Luo and Oyedele, (2021)	LSTM neural network is adopted	LSTM	Educational & Research	Hourly	Cooling	Meteorologic al conditions	2 instances 2.71225 (MAE)
82	Mustaqeem et al., (2021)	proposed an ensemble deep learning-based approach	LSTM	Residential Commercial	Daily Hourly	Total Building Energy		0.36 (MAE)
83	Tian et al., (2021)	four case studies utilized representing diverse energy and time scales	RF recursive feature eliminatio n (RFE) and fuzzy c-means clustering (FCM)	Educational & Research	Hourly Daily Monthly	Total Electricity		4.89 (MAPE)

84	Cho et al., (2019)	Model comparison	SVM RF LR	Commercial	Hourly	Total Building Energy	Meteorologic al conditions	11 instances	8.36 (MAE) 6.97 (MAE) 16.72 (MAE)
85	Jeong et al., (2021)	Proposed a day-ahead electric load forecasting model for buildings	ANN	Commercial Educational & Research	Daily	Total Electricit y		2 instances	6.1945 (MAE)
86	Culaba et al., (2020)	Model comparison	SVM based on reviewed papers	Residential	Hourly	Total Building Energy	Meteorologic al conditions	30 instances	0.0199 (MAE)
87	Ding et al., (2020)	Proposed an evolutionary double attention based.	ANN DNN	Commercial	Hourly	Total Building Energy		1 instance	11.86 (MAE) 24.46 (MAE)
88	Parhizkar et al., (2021)	Proposed pre-processing method to remove noisy features	SVM RF LR	principal component analysis (PCA)	Commercial	Hourly	Total Building Energy	Meteorologic al conditions	1 instance 0.685 (MAE) 0.9925 (MAE) 0.73 (MAE)
89	hwang et al., (2020)	Proposed an advanced energy prediction models using deep learning techniques.	RF GBM	domain knowledg e and correlatio n analysis	Educational & Research	Hourly	Heating Cooling		1 instance 0.23155 (MAE) 0.1288 (MAE)
90	Szul et al., (2020)	Proposed the six most appropriate neural methods	ANN SVM	Residential	Annual	Heating	Occupancy	380 instances 7 (MAE)	7.84 (MAE)
91	Shen et al., (2020)	Developed an improved Support Vector Regression model	SVM	Akaike Information Criterion	Commercial	Daily	Total Building Energy		179 instances 16.698 (MAE)
92	Gassar et al., (2019)	Developed data-driven models	RF LR GBM	Commercial		Total Building Energy Total Electricit y		7200 instances	0.5 (MAE) 3.545 (MAE) 3.96 (MAE)

3.8.1 Error Rate

Statistical or AI based tools have been utilised in various studies to develop energy prediction models (Al-Rakhami et al., 2019; Divina et al., 2019; C. Fan et al., 2017; Khan et al., 2021). In this research, the reviewed articles have utilized machine learning tools (ANN, SVM, RF, XGB, DNN, LSTM, GBM) and statistical tools (ARIMA, LR). Figure 3.8 displays the proportion of studies that utilised the different types of tools. Generally, 23% and 19% of studies employed ANN and SVM respectively, 12% and 3% used LR and ARIMA respectively, among others. Many studies have conducted a comparative analysis of different Statistical, or AI based tools for energy consumption prediction in terms of effectiveness. For example, Kamel et al., (2020) compared XGB and LR; Somu et al., (2020) compared SVM and ARIMA; Kontokosta and Tull, (2017) compared SVM, RF and LR; Li and Yao, (2020) compared ANN, SVM and LR; (C. Li et al., 2017) compared SVM and LR; among others. After the development of an energy prediction model, it is evaluated and compared using performance measures such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Variation

(CV), R-squared (r^2), Mean Absolute Percentage Error (MAPE). However, the most employed performance measures are MAE, RMSE.

Root Mean Squared Error (RMSE) is a performance measure utilized for calculating the variances between predicted value and the actual observed value. RMSE score closer to zero indicates better performance while the higher the error score represents poor performance while Mean Absolute Error (MAE) is a performance measure utilized for the difference between the predicted and the actual observed values at each point in a scatter plot. The best and worst performance regarded the same as RMSE.

The comparison of the various tools is often for purpose of enhancing prediction performance. Overestimation and underestimation of energy consumption will engender an adverse impact on industrial and economic advancements (Somu et al., 2021). Figure 3.8 visualizes a clear comparison between a set of two tools in a chart. The chart consisted of only studies that reported MAE or RMSE which can be clearly quantified and compared based on the lower the error, the better the performance of the model. Table 3.6 reveals the number of studies that reported RMSE and MAE among other performance for each reviewed tool, however only tools of the 72 studies which reported RMSE, and MAE were included in the chart in Figure 3.8. The chart shows the mean average error for a pair of tools in a direct comparison. Few tools show relatively good performance in comparison to some other tools such as ANN, SVM while LR and ARIMA show relatively bad performance, however this is not enough to ascertain an objective conclusion. For a more critical and fair analysis, pie charts in Figure 3.9 were generated to display the percentage of studies that reported that one tool outperforms or resulted in better performance than the other. These charts have matching pairs as Figure 3.8 however Figure 3.9 encompasses all the studies regardless of the performance measure employed. This is because the charts were visualised based on the report on which tool produced better performance with conclusions formed based on different evaluation methods, Hence the pie charts include all 92 studies as shown in the able of reviewed studies or findings studies (Table 3.6). This table shows the performance measures and measures and outcomes of each tool reported in the research.

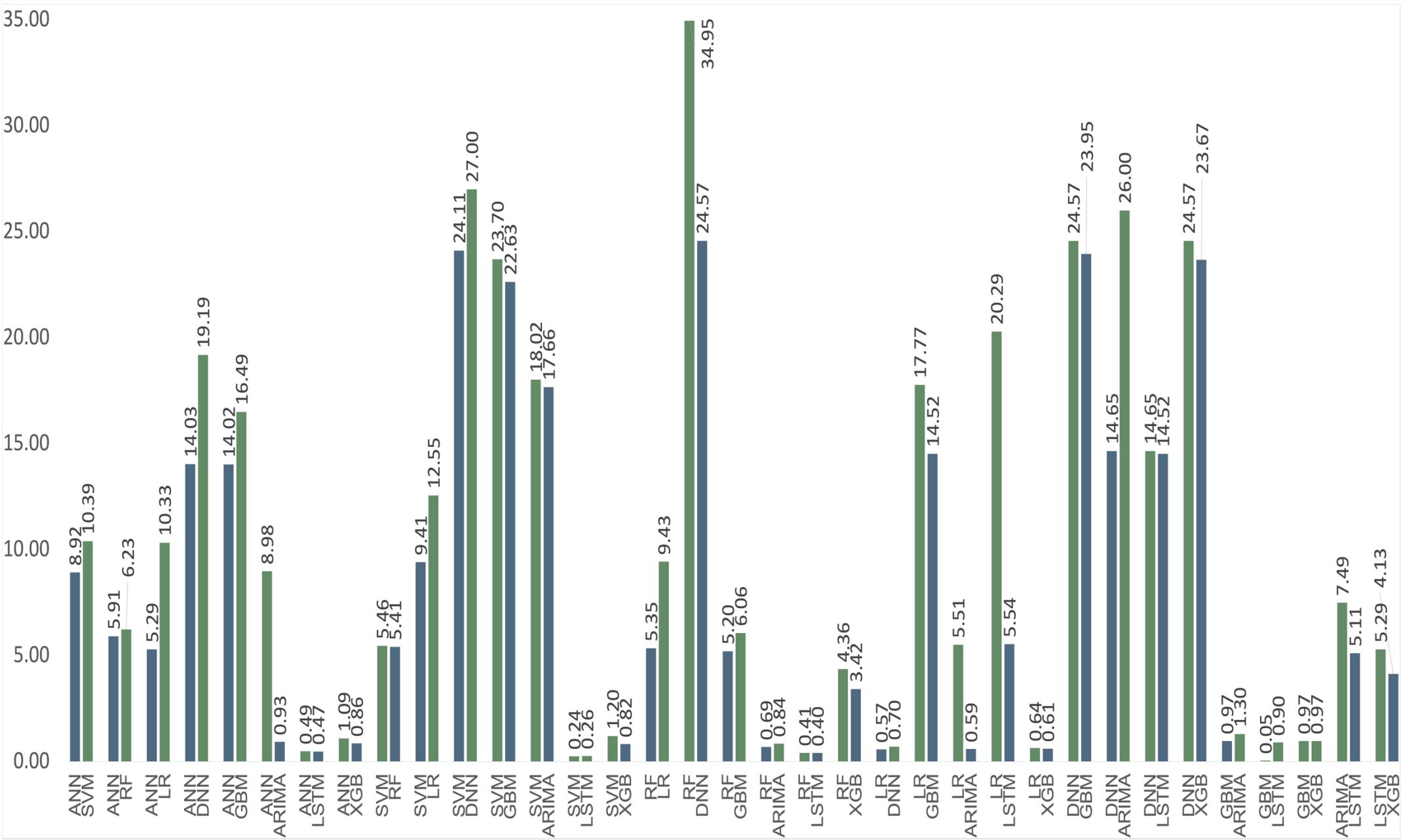


Figure 3.8: Average RMSE and MAE result from studies that conducted a direct comparison of tools

Table 3.7: Summary statistics of the error types reported in reviewed studies

Tools	No. of authors that used tool	No. of authors that reported MAE and RMSE	No. of authors that reported R2, MAPE or CVRMSE
ANN	42	28	14
SVM	52	40	12
RF	31	21	10
LR	28	20	8
DNN	7	4	3
GBM	15	10	5
ARIMA	14	11	3
LSTM	21	16	5
XGB	17	12	5

Table 3.8 shows the number of times a set of two tools were directly compared in the reviewed studies. The majority of the direct comparisons in Figure 3.8, shows that AI based tools produced better performance than statistical tool except in the specific comparisons such as (ANN and ARIMA, SVM and ARIMA). However, Figure 3.9 indicated that more studies reported that ANN produced better performance ARIMA with percentage values of 78% and 22% respectively, while Figure 3.9 also shows that the number of studies that reported that SVM outperforms ARIMA is more than that of ARIMA to SVM with percentage values of 29% and 71% respectively. Figure 3.9 corroborates that AI based tools outperform statistical tools except in Figure 3.9 the comparison of RF and ARIMA, XGB and LR, where an equal percentage of studies specified one is better than the other and vice versa.

Figure 3.8 clearly shows that ANN performs better in the majority of its comparisons, specifically 6 out of 8. The two shortfalls are RF and XGB, which a greater number of studies specified that they produce better than ANN, though, the direct comparison in Figure 3.8 shows that ANN performs slightly better than RF and small variation in performance for XGB and ANN. Also, 5 out of 8 comparisons in Figure 3.8 and Figure 3.9, displayed that more studies infer better performance SVM and GBM respectively. Though, ANN indicates a larger number of studies that reported that ANN produces better performance than both SVM and GBM in Figure 3.9 respectively.

Table 3.8: Matrix of number of times a set of two tools were directly compared in the reviewed studies.

	ANN	SVM	RF	LR	DNN	GBM	ARIM A	LSTM	XGB	Total
ANN	–	25	12	15	6	7	9	10	6	6
SVM	25	–	21	17	4	9	7	6	9	9
RF	12	21	–	11	2	11	2	3	9	9
LR	15	17	11	–	2	6	5	6	6	6
DNN	6	4	2	2	–	1	1	2	1	1
GBM	7	9	11	6	1	–	1	1	3	3
ARIM A	9	7	2	5	1	1	–	9	2	2
LSTM	10	6	3	6	2	1	9	–	3	3
XGB	6	9	9	6	1	3	2	3	–	–
Total	90	73	38	25	5	5	11	3	0	250

Subsequently, further investigation of reviewed studies that employed ANN, SVM or GBM such as (Al-Rakhami et al., 2019; Culaba et al., 2020; Divina et al., 2019; Groß et al., 2021; Li and Yao, 2020; Liu et al., 2021; Shapi et al., 2021; Shen et al., 2020; Szul et al., 2021). Among these studies, Groß et al., 2021 compared ANN and SVM for predicting short term electrical consumption, which reported the ANN produced good performance however SVM produced a slightly better performance with considerable small computational cost. Due to the good performance of SVM, few studies have focused on SVM and conducted an unfair comparison by applying and evaluation different relative kernel functions to investigate the possibility of achieving better performance in comparison with other models (ANN, LR) in its default form such as Li and Yao, 2020. The results of using different SVM kernel function were averaged for a fair and unequivocal comparison. Culaba et al., 2020 found SVM to have superior prediction ability for mixed use buildings in high-dimension datasets. SVM models were developed using the different kernel functions namely Fine Gaussian, Linear Cubic, Medium Gaussian, Coarse Gaussian. Of these functions, Medium Gaussian was considered the most appropriate because it produces the best outcome based on accuracy and training speed. Though, the SVM radial basis function kernel model was applied and compared for predicting long term residential heating and cooling load in the study by Li and Yao, 2020. Radial basis function presented the best performance amongst all the models for the annual residential heating and cooling load intensity prediction. Shapi et al., 2021 explored only radial basis function for energy prediction of smart building, which can be due to its production good result in other studies (Feng and Zhang, 2020; Khosravani et al., 2016; Truong et al., 2021). Few studies have indicates that GBM performed better ANN (Divina et al., 2019; Feng and Zhang,

2020; C. Zhang et al., 2021), GBM has received considerable attention as only 15 studies applied these tools among the 92 reviewed studies. Only one study compared GBM and LSTM, where GBM was reported to be better (Khan et al., 2020). Although Khan et al., 2020 could be said to have conducted a fair comparison, this is not enough to base a firm conclusion. (Groß et al., 2021) employed one evaluation criteria on the basis that performance measures have drawbacks. Considering the drawbacks it is necessary to use multiple evaluation methods in order to create several grounds for conclusion and one can corroborate the other as used in other papers such as (Divina et al., 2018a; Seyedzadeh et al., 2018; Zhong et al., 2019) Figure 3.9 below shows the percentage of studies that reported each tool as better than the other in a pairwise comparison.

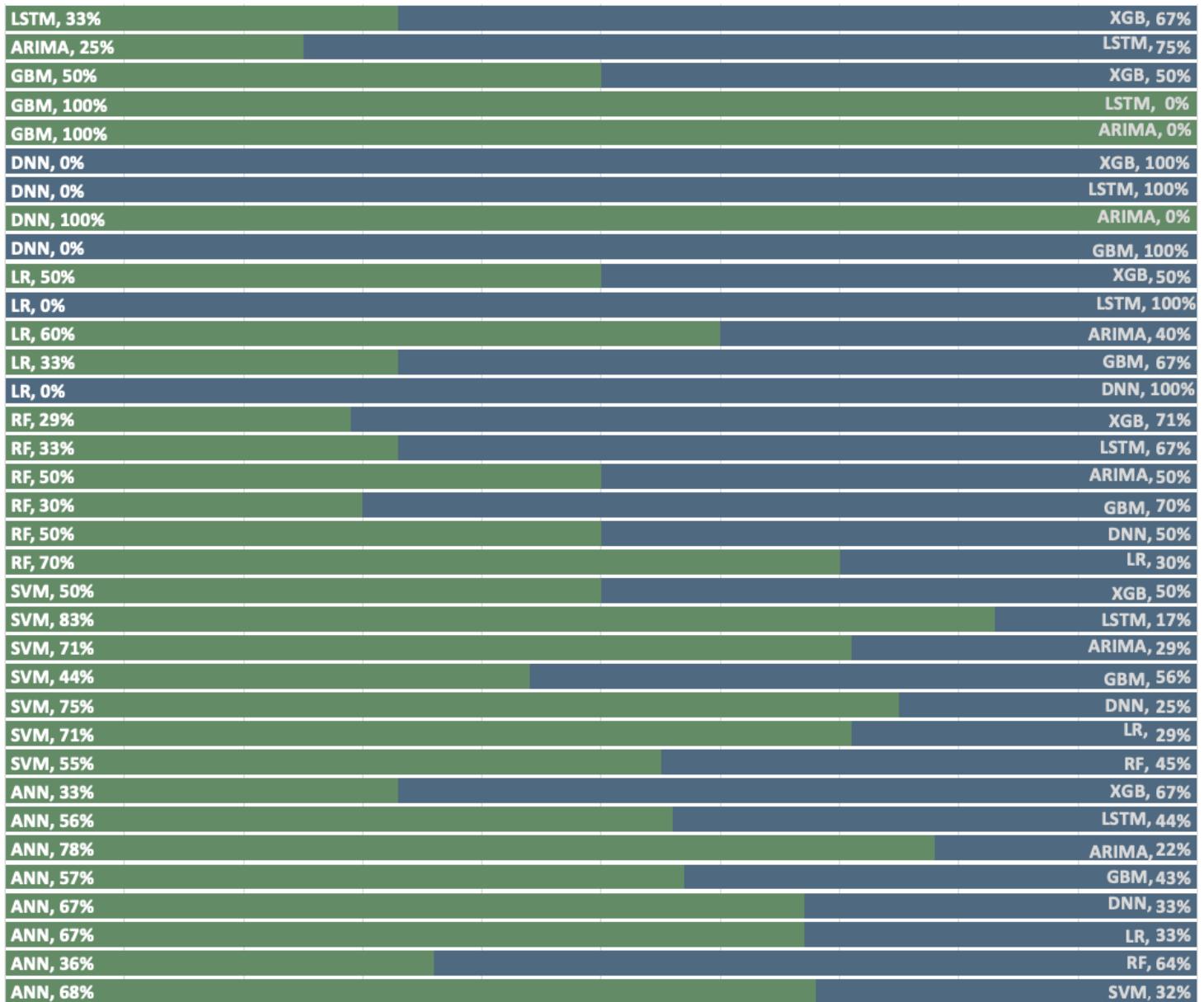


Figure 3.9: Stacked plot indicating the percentage of studies the reported one tool outperforms the other.

3.8.2 Building Type

Buildings of different types have been explored in the development of statistical or AI based energy prediction models. The buildings were categorised into three groups based on their practical usage: Residential (i.e. (Al-Rakhami et al., 2019; Kamel et al., 2020)), Commercial (i.e. Hotel (Ahmad et al., 2018; Borowski and Zwolińska, 2020; Shan et al., 2019), Hospital (Cao et al., 2020; Silvestro et al., 2017)), Educational and research ((Amber et al., 2018; Jang et al., 2019)) buildings. The application and proportion of exploration of different buildings in the reviewed studies are shown in Table 3.6 and Figure 3.7.

The review studies engendered a proportion of 33%, 44% concentrated on energy prediction model development for residential, and commercial buildings, with only 23% concentrated on educational and research buildings. The proportion of studies focused on these buildings is mainly subject to the accessibility and reliability of energy consumption and metadata of these buildings (Fathi et al., 2020b). However, in consideration of non-residential buildings accounting for 67% in reviewed studies, the relatively low proportion of studies concentrated on residential buildings is due to a deficit of sensor-based data which is relied upon for model training. This type of data is easily accessible for non-residential buildings but difficult to obtain for residential buildings as most residential buildings are not metered in a system that permits for sensing at high granularity (Wang and Srinivasan, 2017). Also, the high variation of occupancy behaviour and the inability to easily access occupancy data could be another reason, as occupant behaviour is considered the most uncertain in building energy consumption prediction (Amasyali and El-Gohary, 2018; Khan et al., 2021). Notwithstanding, these challenges need to be overcome because energy consumption prediction for residential buildings is important, as energy consumption for residential space heating and cooling accounts for about 70% in the UK (Li and Yao, 2020) and in the US and EU, about 40% of the energy consumption(Divina et al., 2018a). Hence, further studies need to be conducted on residential buildings. Perhaps, the exploration of adapting or extending commercial building energy prediction models for the residential sector. Furthermore, non-residential building are receiving significant attention due to their high rate of total energy consumption in comparison to residential buildings(Somu et al., 2021). More explicitly, higher education institutions and commercial buildings account for 45% and 30% more energy consumption than residential buildings (Amber et al., 2018).

For building types, the chart in Figure 3.10 displays only studies that reported the MAE and RMSE and the mean average was used for a clear and unambiguous comparison. Figure 3.10

clearly shows that AI based tools produced better performance than statistical tools across all building types except Educational and research buildings. XGB and RF performed better than the other models, while the worst performance was ARIMA for residential buildings. LSTM exhibited the best performance for commercial buildings while GBM indicated the worst performance. ARIMA and LSTM outperform other models for commercial building while the worst performance was SVM. To avoid ambiguity, Ai based or statistical tools that were used in less than two studies were excluded because the mean average will generally be high. Hence, DNN for all building types were not included in Figure 3.10. Likewise, GBM for residential and educational building were not included and LR for educational buildings.

32

27

22

17

12

7

2

-3

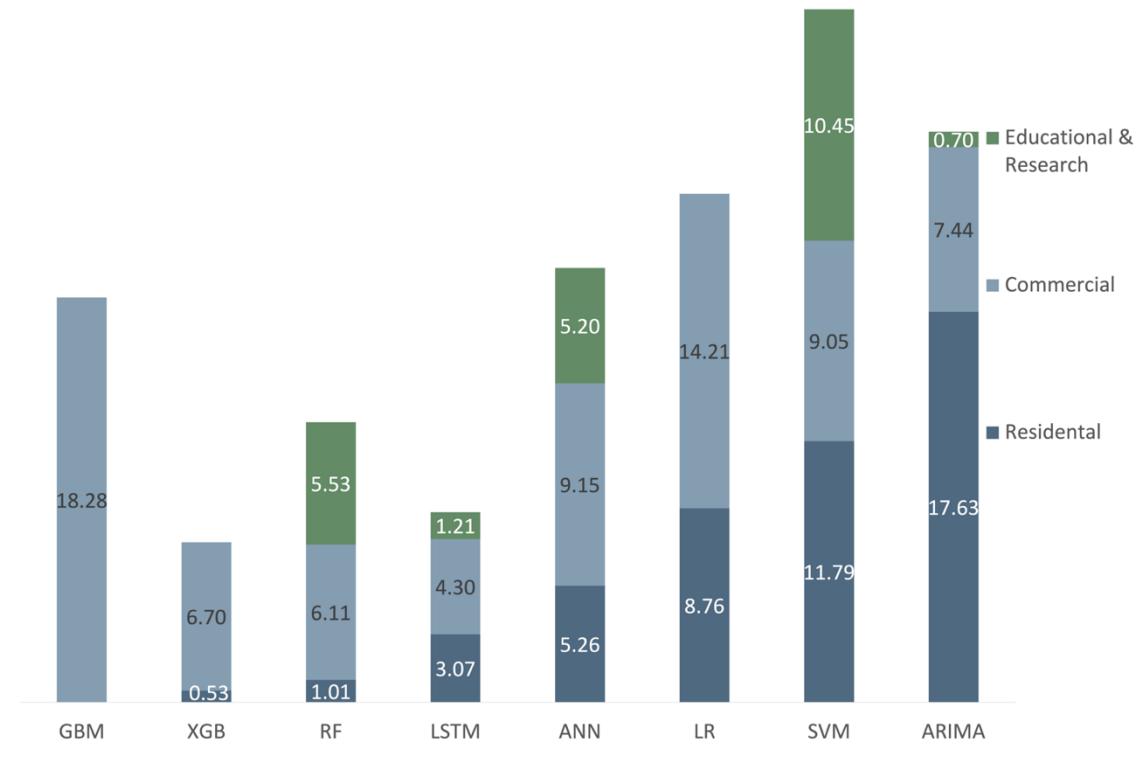


Figure 3.10: Average RMSE or MAE results from studies that specified the building type applied.

Few studies explored the application of certain tools on different building types in the same research(Almalaq and Zhang, 2019; Eseye and Lehtonen, 2020; Groß et al., 2021; Kontokosta and Tull, 2017) as indicated in Table 3.6. 40% of studies compared residential and commercial, 20% used commercial and educational or research building while 13% applied all the building types as shown in Figure 3.11, However 27% of studies did not clearly specify the type of building data used. A further examination of few studies that specified building types, Groß et al., 2021 conducted a comparative analysis of various tools including ANN, SVM, RF among others for daily prediction on the different building types.

The outcome suggest that neural network perform best for non-residential building (e,g. school and supermarket). Groß et al., 2021 stipulated that ANN and LSTM are suitable for data with seasonal pattern and data containing noise. Safa et al., 2017 compared ANN and LR for developing monthly energy prediction of a research office. These tools selected based on its popularity in literature, ANN has been employed more frequently due to its capacity of shaping non linear network(Ahmad et al., 2017b), and LR has been used more frequently than other models such as XGB and GBM as shown in Figure 3.7. Safa et al., 2017 determines that ANN were more proficient at predicting energy consumption of research building than LR. Nevertheless, study by Sadeghi et al., (2020) on the prediction on energy performance for residential buildings using a relatively large data size, demonstrates and emphasizes that DNN outperform ANN. Hence, DNN should receive more attention or be more frequently explore where there is a large sample size.

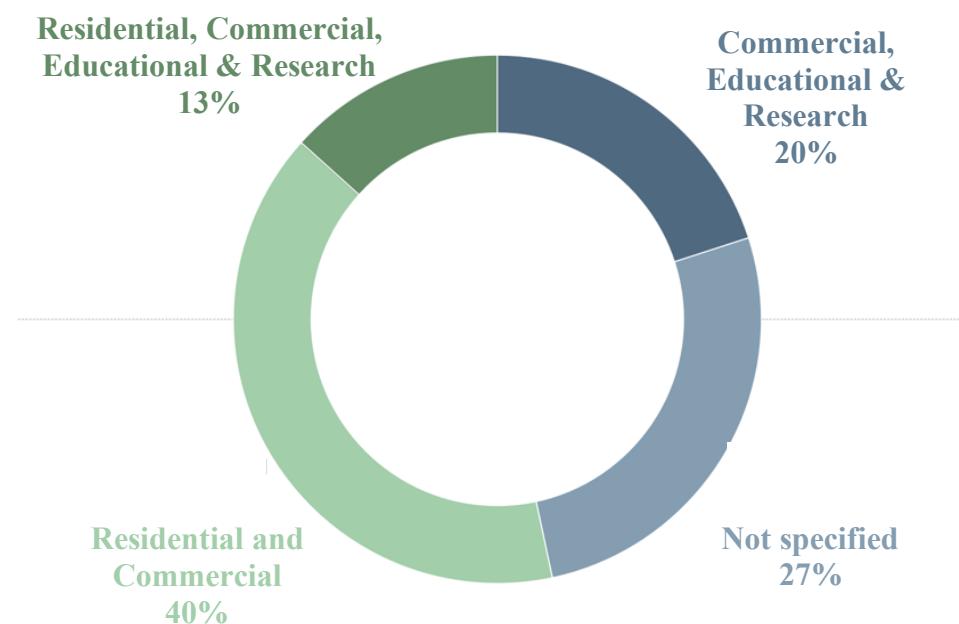


Figure 3.11: Proportion of reviewed studies that used different building types.

3.8.3 Energy Type

Based on reviewed studies, the different types of energy consumed in buildings were classified into five different groups namely Total Building Energy (TBE), Total Electricity (TE), Cooling (CL), Heating (HT), Natural Gas (NG). In this group, TBE and TE represents the sum of all types of energy in the building and the sum of all electrical related energy respectively. As displayed in Figure 3.7c, most studies concentrated on predicting Total Building Energy (26%),

Total Electricity (26%), Cooling (25%), Heating (19%), while only 4% concentrated on predicting Natural gas. This depicts the relative low attention given to Natural gas which can be due to the transition in utilization of renewable energy sources for operational buildings as opposed to non-renewable sources that have gradually been degrading such as natural gas (Fathi et al., 2020b).

TBE receives significant attention because encapsulates other energy types and it accounts for a high proportion of energy consumption. For example, US and EU account for about 40% (Divina et al., 2018a). Following this based on attention received or the percentage of reviewed studies is TE, electricity is the energy source with the most rapid increase in residential and commercial buildings (Y. Liu et al., 2020a). In China, electricity usage in residential buildings consumes about 69.58% and By 2050, it is estimated to account for beyond one-fifth of worldwide building electricity consumption (Capuano, 2019). Subsequently, heating and cooling in residential buildings significantly contribute to energy consumption(Li and Yao, 2020). The development of an accurate prediction model for space heating and cooling load will yield support in the optimizing building properties during the design and retrofitting stage which will progressively enhance energy conservation and reduce carbon emission.

3.8.5 Temporal Granularities

Building energy prediction for different granularities are receiving more attention as buildings have received more attention as buildings are now equipped with smart meters for energy use monitoring at a pulverized interval which assist in comprehending energy consumption patterns for the purpose of energy demand prediction (Somu et al., 2021). The various temporal granularities or time scales utilized in the review studies include yearly, monthly, weekly, daily, sub-hourly energy consumption prediction. Among these, hourly energy consumption is the most explored in reviewed studies with 47% research focus, followed by daily with 23% and sub-hourly with 14% as visualized in Figure 3.7b. These temporal granularities were denoted as short-term energy consumption and their popularity in review studies is due to its direct association to the daily building operations (Fan et al., 2014b). However, only 8% of the reviewed studies each concentrated on model development for prediction of monthly and yearly energy consumption. This could the subject to the requirement of a considerable amount of data which accounts for long time duration in order to generate good result (Amasyali and El-Gohary, 2018). Also, the nonlinearity in long-term sample is often more pronounced in comparison to short term data (Li et al., 2015). Despite the issues encompassed with the

development long-term energy consumption prediction models, it is considered very essential as it helps in making informed decisions for financial and operational planning (Lee and Rhee, 2021).

Reviewed studies have reported specific tools to have produced good performance on different granularities of energy prediction. For example, (Liao et al., 2020) reports that RF yields the best prediction performance for monthly electricity consumption prediction. (Li and Yao, 2020) shows that SVM model using the gaussian radial basis function kernel exhibited the best prediction performance for the prediction of annual residential space heating and cooling load. Whereas Cao et al., (2020) found RF, XGB, and SVM were the most accurate ensemble models for daily electrical energy prediction. (Moon Keun Kim et al., 2020) stated that both the LR and ANN models were able to satisfy the requirements for long-term and hourly electricity usage prediction for a building, however, case dependent on the occupancy degrees and local environmental conditions.

3.8.6 Data Size

It is well noted that the size of data is important in developing energy prediction models (Liang Zhang et al., 2021). AI based tools rely on historical data to learn energy consumption patterns (Somu et al., 2021). However, some tools tend to thrive better in small datasets such as SVM (Mat Daut et al., 2017) while some require large datasets to produce good results such as DNN (Ngarambe et al., 2020). Feng and Zhang, (2020) emphasized the consensus that the simple application of a single model for prediction is not suitable as the performance of a prediction model is considerably influenced by the data and no single model in its stand-alone form can be noted the best amongst all models. Table 3.6 shows the reviewed studies that utilized different sizes of data, 40% of reviewed studies explored the utilization of more than one building data.

DNN is recognized for producing good performance in a large dataset and this has been analysed and performance was emphasized in the study by Sadeghi et al., 2020. (Amber et al., 2018) stipulated that DNN models are not as favourable in studies with limited amount of data thus its performance relies heavily on a large amount of data. Due to the small amount of data with high dimension used in the study by Guo et al., (2020b), RF produced the worst tools while a hybrid model, boruta feature selection applied with SVM produced the best prediction performance among others.

3.8.7 Feature Selection

Feature selection plays a significant role in improving the performance of a prediction model, by eradicating the unimportant and irrelevant noisy feature to further enhance the quality of the dataset (Asir et al., 2016). Feature selection is not well addressed in energy use prediction literature (Hsu, 2015). Hence, several studies selected variables based on academic literature or domain knowledge(Cao et al., 2020; Y. Liu et al., 2020a; Somu et al., 2020). Table 3.6 shows the number of studies that employed a feature selection method while the remaining studies selected features based on popularity, domain knowledge or academic literature.

Some studies have buttressed the need for feature selection to improve the accuracy of prediction model, for example (Liu et al., 2021)stated that the most important input features should be selected when developing an energy prediction model, especially using ANN. (Guo et al., 2020b) conducted a comparative analysis of three feature selection methods namely stepwise, least angle regression (Lars), and Boruta. Of these three, boruta emerged the best and it was stated that a hybrid model (SVM and Boruta) produces the best result among other models (i.e., GB and RF). Hosseini and Fard, (2021) employed univariate regression algorithm feature selection method and concluded RF as the best tool. Additionally, feature selection methods, such as the mutual information-based filter techniques identify features centred on dependencies of the dependent or target features. However, filter techniques rely on statistical connections, which do not consider the prediction process. Consequently, there is a need to employ wrapper methods, which select features based on their influence or effect on the prediction performance (Feng and Zhang, 2020).

3.8.8 Proposed Framework

Drawn from the outcome and findings of this review, a simplified framework was created to aid BEP model developers in making the right decision, when selecting the prediction tools based on the given situation (e.g., data availability, building type etc.) as shown in Figure 3.12 below. Fundamentally, all tools employed in the development of BEP models can productively make predictions. Nonetheless, certain tools are more powerful than others in certain situations.

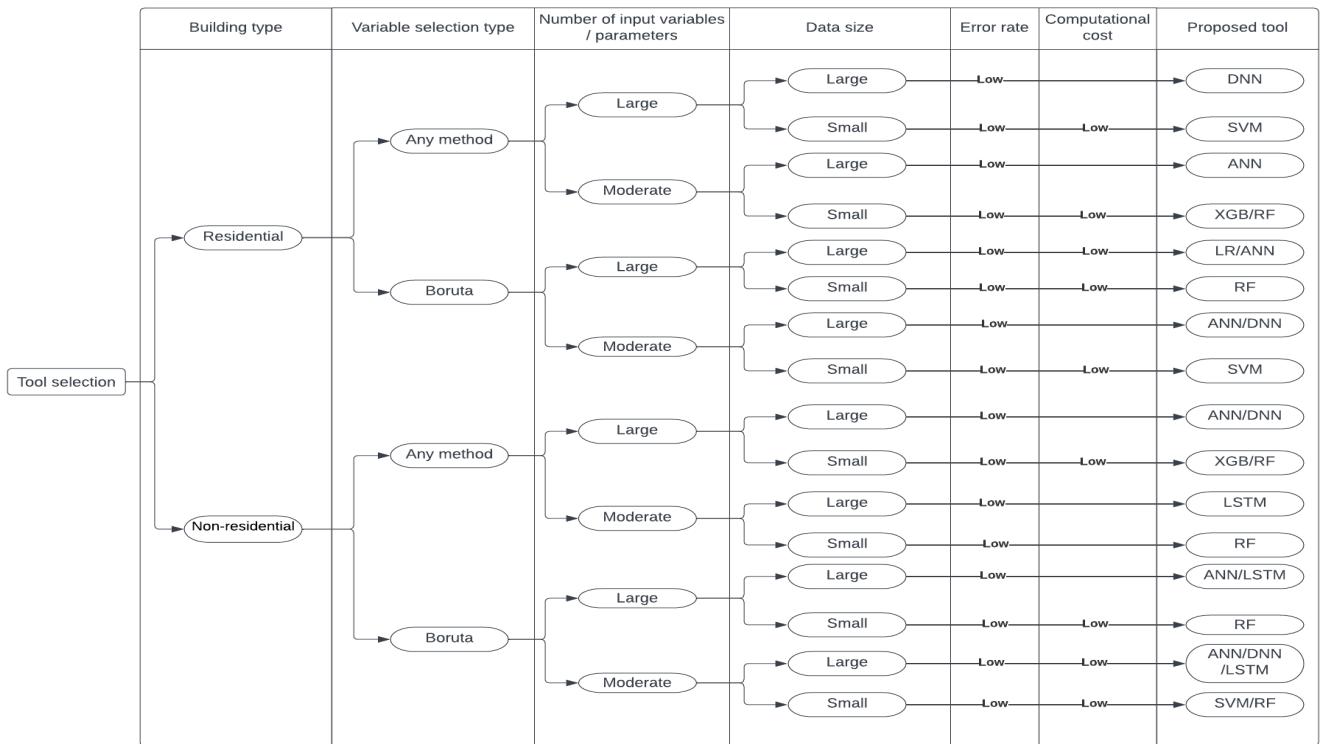


Figure 3.12: A simplified framework for tool selection in diverse situations

The framework explicitly depicts that the achievement of the best performance of a BEP model is determined by the appropriate selection of tools based on output type and characteristics of available data. This framework will ensure that tools are not selected based on popularity and unacademic factors. Also, BEP model developers will be able to select tools based on requirements. For instance, if a low error rate is considered a top priority based on clients' specifications but the data available is small. ANN will not be wrongly selected rather, using this framework SVM will be selected as the appropriate tool for such a situation.

The implication of the utilization of this framework in practice is that it will engender the optimal usage of tools for the right situation based on their strengths and incite a more efficient method of BEP model development bespoke to clients' preferences. As this will ameliorate the time-consuming method of developing and comparing many BEP models using various tools to select the most suitable for the given situation. Additionally, the implication of research is that it will serve as a guideline for researchers when choosing the most suitable tool for the specific situation and aid in alleviating the random or popularity-based selection of tools.



Figure 3.13: Strength and weakness of tools based on review

Based on the results and deductions of this systematic review, Figure 3.13 presents a streamlined framework to support or guide the selection of tools by energy prediction researchers and model developers. Essentially, all of the tools reviewed in this research could make predictions. However, some tools are better than others in specific situations. For example, if a low error rate is the target goal for model developers and only a small sample size is available for model training, SVM will be a suitable choice for such a situation.

In Figure 3.13, the singular rectangles around the circle show the strengths and weaknesses of each data-driven tool, and their ability to produce good predictions in certain conditions. For example, ANNs are noted to be computationally expensive; however, they produce good predictions for the energy use of residential and commercial buildings.

3.8.9 Theoretical and Practical Implications

Over the past few years, the selection of tools for building energy prediction has typically been done arbitrarily or based solely on their popularity, without taking into account their strengths and weaknesses in certain conditions (Divina et al., 2018b; Feng and Zhang, 2020; D.M.F. Izidio et al., 2021; C. Robinson et al., 2017). This approach has engendered poor performance and time-consuming comparative analysis of tools in studies. Despite the prevalence of these efforts, well-informed tool selection for a certain condition can result in the development of more accurate prediction models and more efficient comparative analysis of tools in studies. This research conducted a systematic literature review of popular and favourable building energy consumption prediction tools and based on the findings, conducted a comparative analysis of the identified higher-performing tools on a standard dataset in a specific condition. This will lead to well-informed utilization of tools for building energy consumption prediction.

Regarding theoretical implications, by analysing the performance of statistical and AI-based tools across different conditions and identifying their strengths and limitations, this research contributes to the development of a better understanding of the effectiveness of these tools in different scenarios. Different developers attempt to develop models with a goal in mind, however, the achievement of this goal is predicated on the conditions. For example, accuracy is of great importance when developing a BEP model however the accuracy rate is highly correlated to the input/output, and data size amongst other conditions(Fathi et al., 2020b; Goyal et al., 2020; Runge and Zmeureanu, 2019). The research provides insights into the factors that

influence the performance of these tools, which can inform the development of new hybrid models and improve the accuracy of building energy prediction models.

Furthermore, this research has significant practical implications, as this can inform the development of new building energy management systems that leverage the most effective statistical and AI-based tools. It can inform decision-making related to the selection and implementation of building energy management systems. Specifically, the development of a more accurate model through the identification of the most effective statistical and AI-based tools can help practitioners and researchers in the field of building energy management optimize energy consumption and reduce costs, leading to improved sustainability and reduced environmental impact.

Moreover, the comparative analysis conducted using a standard dataset will provide some form of assurance to practitioners and researchers. Hence, this research can inherently lead to the development of new energy prediction models that are tailored to specific conditions such as different types of buildings and energy sources, among others. By identifying gaps in the literature and highlighting areas for further research, the research can also inform future research efforts in the field of building energy management.

3.9 Chapter Summary

This chapter presents a systematic exploration of existing literature and theoretical frameworks related to building energy prediction tools. The research methodology section demonstrates the method employed, including data collection methods, search criteria, and inclusion/exclusion criteria. A bibliometric analysis is then conducted, examining publication trends, keyword co-occurrences, top journals, and global collaboration patterns. Subsequently, a systematic analysis is performed, evaluating statistical and AI-based tools based on relevant criteria. The results and discussions section covers various aspects including error rates, building types, energy types, input features, temporal granularities, data sizes, and feature selection. Overall, the result indicates that no singular tool is primarily better than all other tools across the identified criteria and emphasizes that no one tool is suitable for all purposes under different circumstances. Thus, it is evident that All tools have their strengths and drawbacks and produced different outcomes under different circumstances (i.e., data conditions, developer goal, among others). Finally, the chapter

concludes with the presentation of a theoretical framework and discusses the theoretical and practical implications of the findings.

CHAPTER FOUR

4.0 SYSTEMATIC LITERATURE REVIEW: ANALYSIS OF FEATURES INFLUENCING BUILDING ENERGY PERFORMANCE

4.1 Chapter Overview

In this chapter, a comprehensive and systematic review of literature is conducted to identify the most common and relevant features that influence energy consumption of buildings. The increasing need to decrease the intensification of energy in buildings has made it imperative to understand the features influencing energy performance in building (EPB). This identification of the key features that affect EPB will help inform the decision of researchers and practitioners during the feature selection process for development of high performing BEP model. Furthermore, this will avail architect or building designer with the knowledge to support decision making for the curation of energy efficient designs (Suh and Chang, 2014). This chapter systematically analysed journal articles that have explored the various features influencing energy use in buildings. Furthermore, this chapter conducts a quantitative bibliometric analysis to pinpoint the trends and examine knowledge gaps. Also, this chapter helps to achieve Objective #1 of this research.

4.2 Significance of Feature Analysis

Energy efficiency is considered a viable solution to the ubiquitous problem of high energy consumption in buildings, consequently leading to significant increase in carbon footprint, which is ecologically detrimental (Allouhi et al., 2015; Bhattacharjee and Reichard, 2011). The estimation and optimization of building energy performance (BEP) is one of the most efficient approaches for building energy efficiency. However, there are various features influencing building energy performance such as building envelope(B. Chen et al., 2021; Liu et al., 2022), climate(Košir et al., 2018; Mafimisebi et al., 2018; Park et al., 2020), occupant behaviour(Laaroussi et al., 2020; Wu et al., 2020) among others (Ciulla et al., 2016; Laaroussi et al., 2020). It remains a prerequisite that, the enhancement of energy efficiency in the building sector is predicated on the ability to clearly comprehend current features affecting energy consumption (Liu et al., 2022). As stated by Lin et al., (2018), Irrespective of the method

employed to improve building energy efficiency, whether applying an energy policy regulation, developing building energy prediction model (BEPM) or creating a green certification program, an essential pre-requisite is the comprehension of the features influencing energy consumed in the building stock. BEPM is measured as the most promising solution for improving energy efficiency (Khan et al., 2021; Olu-Ajayi et al., 2022b; G. Zhang et al., 2020). Researchers have developed various building energy prediction models (BEPM) using several features such as building envelope(i.e., wall, roof) among others(Khan et al., 2021; Kim and Suh, 2021; Mocanu et al., 2016). However, most of these features are selected based on popularity in research, which often leads to the generation of inaccurate results.

According to the International Energy Agency (IEA), energy performance in buildings is influenced by technical or physical features (i.e., climate, building envelope and equipment) and human features (i.e., operation and maintenance, interior conditions and occupant behavior) (Yoshino et al., 2017). Subsequently, many studies have underlined these features as important elements for understanding energy usage in the buildings(Mamdooh Alwetaishi and Benjeddou, 2021; Huang et al., 2013; Ihara et al., 2015; Iken et al., 2019; L. Y. Zhang et al., 2017). Although it is a growing consensus that operation and maintenance configurations are key features of BEP (Laaroussi et al., 2020). Several studies stipulated that physical features such as building envelope constitute the most significant effect on BEP(Košir et al., 2018; Ocampo Batlle et al., 2020; Paukštys et al., 2021; Su et al., 2021). Also, numerous studies have explored the effect of climate change on BEP, such as the effect of changes in temperature(Ciulla et al., 2016; Lee and Lee, 2009; Mafimisebi et al., 2018; Meng et al., 2020). The International Energy Agency stipulated that due to the increase in temperatures during summer, the demand for cooling in buildings escalated abruptly, leading to significant increase in energy consumption("International Energy Agency: Cooling," 2019).

In recent years, there has been increasing research on several components of the building envelope that can significantly improve building energy efficiency(Laaroussi et al., 2020), For example, (Ocampo Batlle et al., 2020) stated that the better insulation of windows, walls and roofs among others can contribute to efficient usage of energy. Nevertheless a few studies(e.g., Lin et al., 2018; Yang et al., 2021) still emphasised the importance of operation and maintenance. For example, (Lin et al., 2018) concluded that Air Condition (AC) cleanliness and proper housekeeping is a critical factor to ensure energy savings. Regarding performance and energy efficiency, other design features such as windows are identified as the weakest

component of the building envelope, accountable for a significant amount of heat transmittance and thermal bridging in buildings (Mamdooh Alwetaishi and Benjeddou, 2021).

In the field of building energy efficiency, thermal building insulation is the most researched area, as insulation materials have been identified as a successful method to minimize energy use and therefore aid the achievement of sustainable buildings (Iken et al., 2019; William et al., 2021). (Li et al., 2014) stated that a building formation with fixed envelope properties have been acknowledged as a driver affecting the heating and cooling loads in residential buildings in certain cities such as Rome and Hong Kong. Depending on the weather conditions in various cities and the appropriate optimization of building envelope, features such as windows can save over 25% of total heating load (M. Alwetaishi and Benjeddou, 2021). Despite the highlighted increase in building insulation, (Suh and Chang, 2014) argue that the changing the orientation of the building will impact building energy performance better increase or decrease in wall insulation. This could be said to be location dependent considering external wall insulation of building in cold areas in China, have achieve remarkable energy saving (L. Y. Zhang et al., 2017).

Apart from a few studies(M. Alwetaishi and Benjeddou, 2021; Rusek et al., 2022), most studies are concerned with the understanding these features affecting existing buildings and this can be subject to the fact that most of the existing buildings were constructed before energy efficiency in building was a concern, and many of these buildings will remain functional beyond 2050 (Huang et al., 2013). However, few studies emphasized the importance of considering energy efficiency at the early design stage of buildings (Fan, 2022), as the decisions made in the early design stage of building (i.e building components selection, orientation among others) can tremendously reduce or increase building energy consumption(Suh and Chang, 2014).

Several studies conducted evaluation of specific features based on its popularity or domain knowledge. Some studies (such as (Andargie et al., 2019; Paone and Bacher, 2018)) focused on a review of occupant behaviour while some studies(Li et al., 2021; Sadineni et al., 2011) focused on building envelope and environmental features based on their popularity in research. The importance of understanding all the features influencing energy consumption cannot be overemphasized. This is because regardless of the method selected to improve building energy efficiency, there is the need to conduct a comprehensive evaluation of building energy

influencing features which can be useful in consideration of retrofitting existing building and at the early design stage of buildings. This will avail designers with the knowledge to aid them in making optimum decisions for energy efficient design in the early design stage of building. Also, this will aid developers in making appropriate selection of features for the development of estimation models or BEPM. Thus, this chapter delivers a holistic, structured and comprehensive review of studies that have explored various features affecting energy use in buildings and produce a comprehensive theoretical guideline for selecting the most important features for the development of estimation models and likewise provide more knowledge for the appropriate selection of features to optimize when constructing energy efficient building designs.

This research will convey a plausible contribution to knowledge by presenting BEP features needed to develop a high performing BEPM, as exclusion of specific features can easily engender poor BEPM. This will significantly improve the BEPM development process more efficiently. The extent of this review is limited to identification of the most important features for BEPM development. The scope of this chapter is structured as follows: the next section explains the systematic review methodology utilized in this research. Subsequently a bibliometric analysis is conducted and the result and finding described in the next section. This is followed by data analysis section, explaining the steps conducted in analysing the data from the systematic review. The next section presents the result and findings of the analysis. Section six gives the discussion and proposed model while the next section concludes the work.

4.3 Methodology

This research employed a systematic review method to identify the most important features needed to develop high performing BEPM. While some features have been recognised as noncrucial features of energy usage in buildings(Li et al., 2014), There are still so many features that have been considered to affect energy consumption. As a result, the most prevalent features as noted by Yoshino et al., 2017 in an empirical analysis of BEP were classified as technical features (i.e., climate and building envelope) and human features (i.e., interior conditions and occupant behaviour). This review covered the most prevalent and popular features namely building envelope (building floor area,(Li et al., 2014; Liu et al., 2021), window(H. Chen et al., 2021; Suh and Chang, 2014; Yun and Steemers, 2011), roof (Chiradeja and Ngaopitakkul, 2019; Košir et al., 2018), wall (Huang et al., 2013; Ihara et al., 2015; L. Y. Zhang et al., 2017),

orientation (Ghosh and Neogi, 2018; Suh and Chang, 2014)), climate conditions (Rouleau et al., 2018; S. K. Verma et al., 2022; William et al., 2021) and occupancy (Laaroussi et al., 2020; Wu et al., 2020; Zhu et al., 2021).

Among these features, Occupancy is considered to be the most important driver for improving BEP (Zhu et al., 2021). As stated by (Yang et al., 2021) “Buildings do not consume energy, people do”. However, other studies argue that climate condition such as temperature is a more important, as it influences the energy used by occupants/people (Košir et al., 2018; Peng et al., 2021). This research conducts a systematic review of the aforementioned features and a bibliometric analysis of features that improve BEP. Primarily, each driver is accumulated using a systematic literature review and classified or ranked based on its occurrence or frequency of usage.

Systematic review is an academically accepted approach for engendering valid and reliable contribution as it reduces bias thus its popularity in diverse important research fields around the world (Schlosser, 2007). A systematic literature review should follow to a specific process to elicit repeatability, transparency, and rigour. Bibliometrics analysis is a well known and reliable method of analysis for evaluating the development and quality of research produced(Krishnamoorthy et al., 2009)

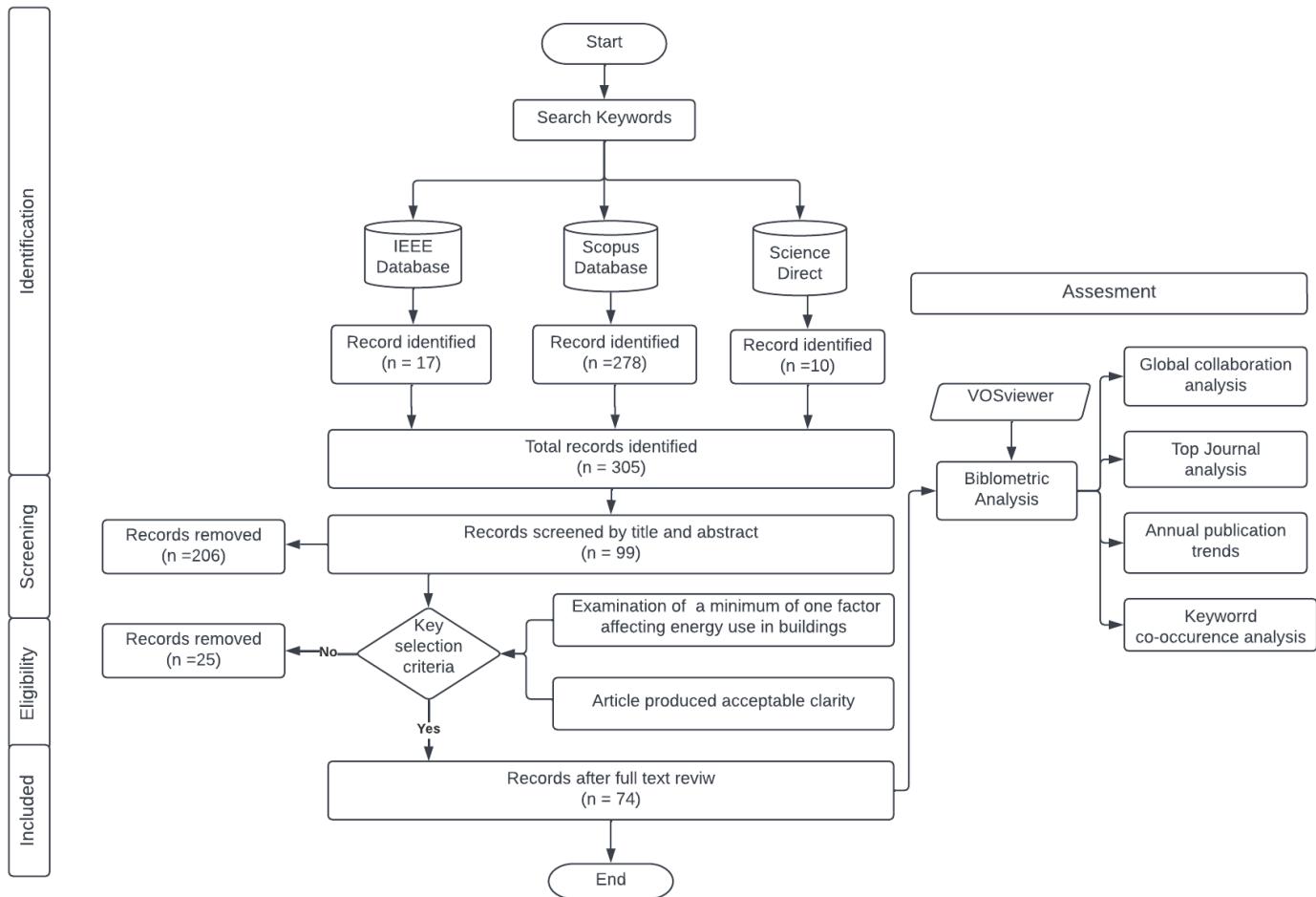


Figure 4.1.: Framework of the key phases of the methodology

4.3.1 Data Collection

This research explored three different databases namely Scopus, ScienceDirect and Institute of Electrical and Electronics Engineers (IEEE) for research articles relevant for identifying the features affecting BEP. Although Other databases were also considered such as Engineering Village (EV), Web of Science, Google Scholar among others based on their popularity and recognition for publishing high quality journal articles, they were not explored due to inaccessibility restrictions and other (Google scholar) due to its elicitation of infinite outcomes with varying accuracy from expected results as corroborated by (Falagas et al., 2008). The databases employed have been considered suitable and sufficient for a systematic review based on its high indexing rate and extensive publication coverage (Debrah et al., 2022; Diir and Santos, 2019). These databases were also selected because they are mutual amongst Q1 energy journals such as Journal of Building Engineering and Energy for Sustainable Development, among others and the utilization of these three database eliminates databases and geographic

bias as covers articles from various countries worldwide, and consequently ensures high reliability and quality(Schlosser, 2007). (see Figure 4.1).

The terms and keywords used for the databases were carefully selected based on review of other energy related papers(Chiradeja and Ngaopitakkul, 2019; Rouleau et al., 2018). These terms and keywords searched across the three databases (i.e., Scopus, ScienceDirect, IEEE) are "building energy consumption" or "building energy performance" or "building energy savings", "features" or "features" as displayed in Table 4.1.

Table 4.1: Database, terms/keywords and research articles search outcome

Databases	Query String	Results
Scopus	TITLE-ABS-KEY (("building energy consumption" OR "building energy performance" OR "building energy savings") AND ("factors" OR "features")) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT- TO (SUBJAREA , "ENER")) AND (LIMIT-TO (LANGUAGE , "English"))	278
ScienceDirect	TITLE-ABS-KEY (("building energy consumption" OR "building energy performance" OR "building energy savings") AND ("factors" OR "features")) AND (LIMIT-TO (SRCTYPE , "j")) AND (LIMIT- TO (SUBJAREA , "ENER")) AND (LIMIT-TO (LANGUAGE , "English"))	10
IEEE	("Abstract":"building energy savings") OR ("Abstract": "building energy consumption") OR ("Abstract":"building energy performance") AND ("Abstract":"features") AND ("Abstract":"factors") Filters Applied: Journals	17
Results identified after full text review		74

There were no restrictions in the search centred on language, document year, as this research considered all research articles from inception that focused on relevant features affecting BEP and all articles generated from the search were written in English. The search results produced articles from 2000 to 2022. The titles and abstracts of the search results were examined to confirm the suitability of the articles for this research. One of the inclusion criteria for selection was that the BEP research must be centred solely, or mainly on a driver or features affecting BEP. Other criteria includes the article must be comprehensive and produce acceptable clarity (i.e., proper elucidation of methodology and conclusions). The abstract and titles of each article were typically sufficient enough to determine the studies that were most suitable for this study;

else introductions and/or conclusions of the article was examined. In some cases, the full text of the article was examined.

To enhance the validity of this research, one exclusion criteria was the restriction of the results to only journal articles, mainly because they are measured as more credible (Schlosser, 2007). Another exclusion criteria is the dismissal of non-English articles, due to the lack of wherewithal to cover interpretation cost. Therefore, 5 non-English articles were removed from the search records. one example of the articles removed due on language constraint is (Chen et al., 2006) which is scripted in Chinese. 36 review articles were also removed as they will primarily comprise of features identified from other research articles. Furthermore, the search result generated articles that were not within the scope of this research, articles from diverse subject areas such as chemical engineering(A. Verma et al., 2022), econometrics and finance(Manfren et al., 2020; Zaidan et al., 2021) among others. The occurrence of those articles could be due to the use of certain term/keyword in search query (i.e., “energy savings”, “features” etc) in abstract or title of article, Thus the subject area was limited to energy related areas only. After the application of inclusion and the exclusion criteria, the final result totalled 74 articles which were reviewed in this research. Thereafter, the bibliographic data of the relevant articles were exported from all databases (Scopus, Science and IEEE) for analysis, before further amalgamation of the data for this research.

4.4 Bibliometric Analysis

Bibliometric analysis was implemented to understand and assess knowledge areas. Therefore, as necessary, various tools were explored and the most appropriate tool was selected (Darko et al., 2020). There are various bibliometric analysis tools such as VOSviewer® (Van Eck and Waltman, 2020), CiteSpace® (Chen, 2014), Gephi® (Cherven, 2015). However, VOSviewer® is the most generally exploited in academic research (Debrah et al., 2022; Li et al., 2021) and it is recognised for its user friendliness and noted as the best tools for bibliometric analysis (Boopathi and Gomathi, 2019; Saka and Chan, 2019), Therefore it was selected and utilised in this research. VOSviewer® is an tool that provides the essential features needed for science mapping and analysis of bibliometric networks (Darko et al., 2020). In this research, the bibliometric analysis was implemented for publication trend analysis, keyword occurrence analysis and Geographical/co-authorship and citation analysis.

4.4.1 Publication Trends Analysis

Generally, the number of publications analysing relevant features in the BEP field from 2000 to 2022 shows increase from certain years (Figure 4.2). The first 13 years (2000-2013) can be said to have received no significant attention or development. However, it is noted that the first few publications within these years were from east Asia (i.e., China (Huang et al., 2013; Lu et al., 2013), South Korea(Yun and Steemers, 2011)), except (Florides et al., 2002) which emanated from Cyprus. This early investigation in China is most likely due to the extreme energy consumption in buildings(Somu et al., 2020). The stable and slow growth in publications across other countries such as United States(Li et al., 2014; Min et al., 2015; Wu et al., 2020), Brazil(Geraldi and Ghisi, 2022; Melo et al., 2014) among others, started from the year 2014 . This phenomenon could be subject to the availability of the worldwide prominent climate reports, such as Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Synthesis Report presenting imminent need for reduction of greenhouse gas emission from buildings, stating the pervasive impacts and the need to build a more sustainable future for posterity(“AR5 Synthesis Report: Climate Change — IPCC,” 2014)

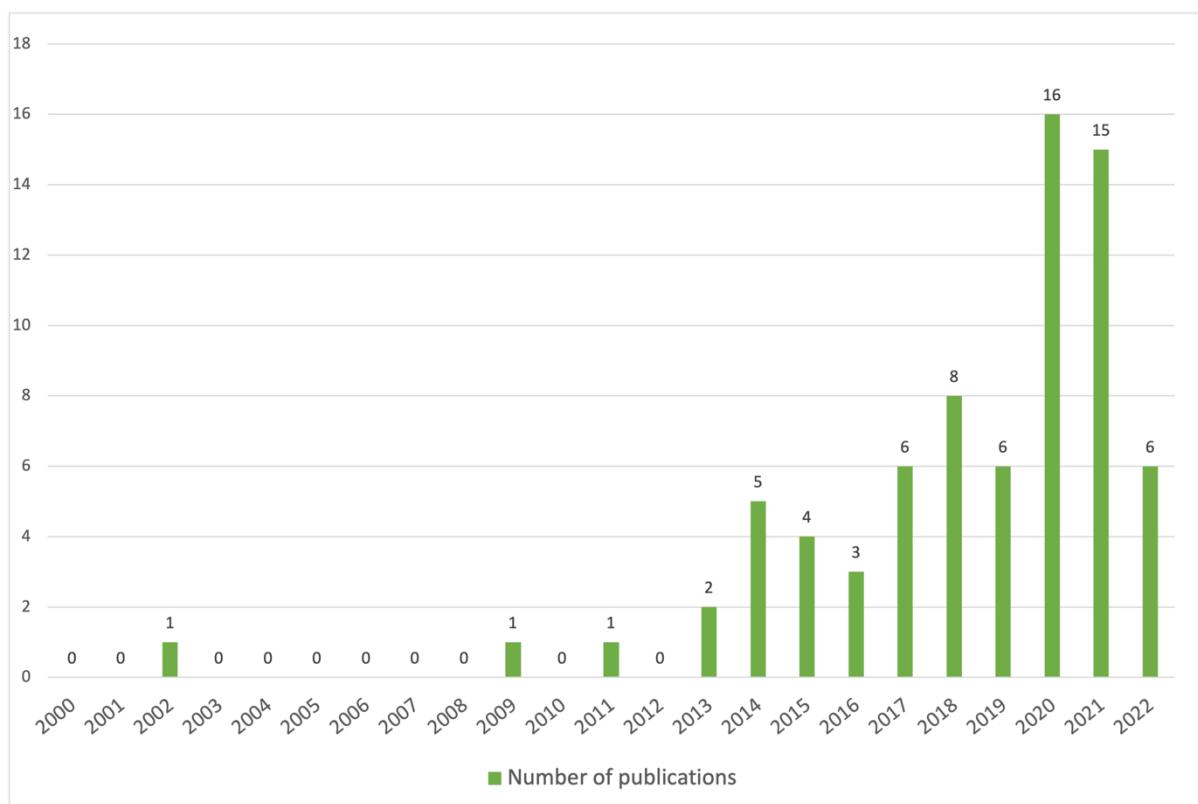


Figure 4.2: Proportion of annual publication on BEP features

4.4.2 Keywords Co-Occurrence Analysis

The examination of certain keywords is often supported for connecting of key research areas in academic literature (Yin et al., 2019). The keywords occurrence map was visualised using VOSviewer. The VOSviewer generates a bibliographic map centred on distance, where the relational strength is denoted by the distance between two keywords, and a short distance signifies a more solid relationship(van Eck and Waltman, 2014). The labels size signifies the degree of the keywords in pertinent studies. The articles produced 876 keywords gathered by utilizing fractional counting. The minimum number of occurrences for each keyword was set 5, therefore only 37 met the threshold as displayed in Figure 4.3.

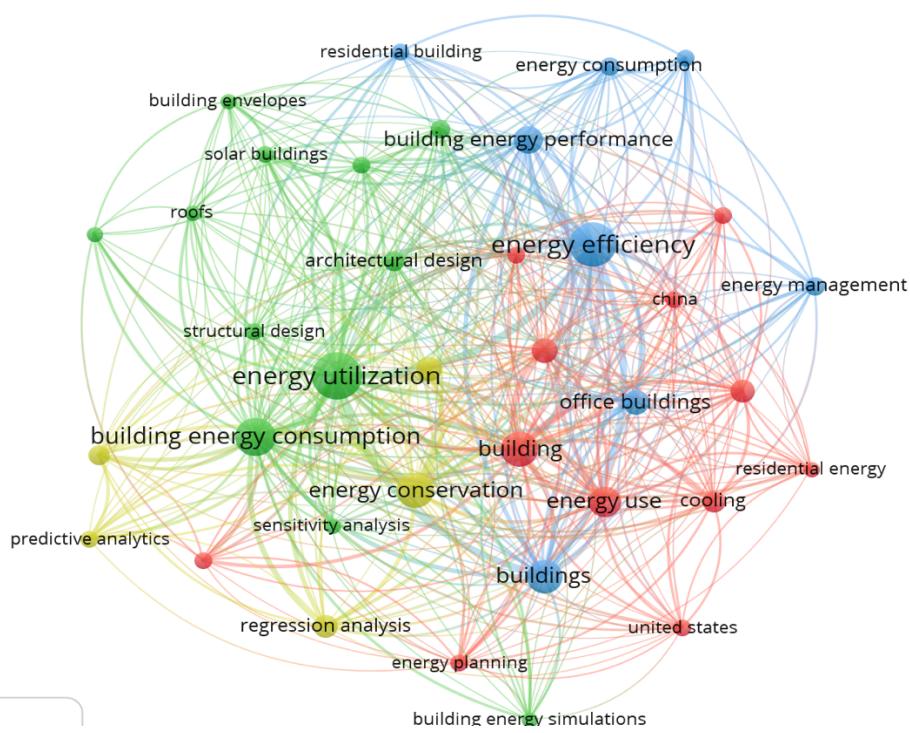


Figure 4.3: keywords occurrence bibliographic map

Based on the highest to lowest number of occurrences, top 20 keywords were selected and visualized in Figure 4.4. This result shows that certain fields in research or keywords are receiving significant attention while some are not receiving as much. For example, “Energy utilization”, “Energy efficiency”, “Building energy consumption”, “Energy conservation”, “Building”, “Energy use” among others, have been abundant in BEP research and this also substantiates the increase in research on improving building energy efficiency. Walls and architectural design were also noted in the top 20 keywords which depicts that these features are receiving attention with respect to improving BEP.

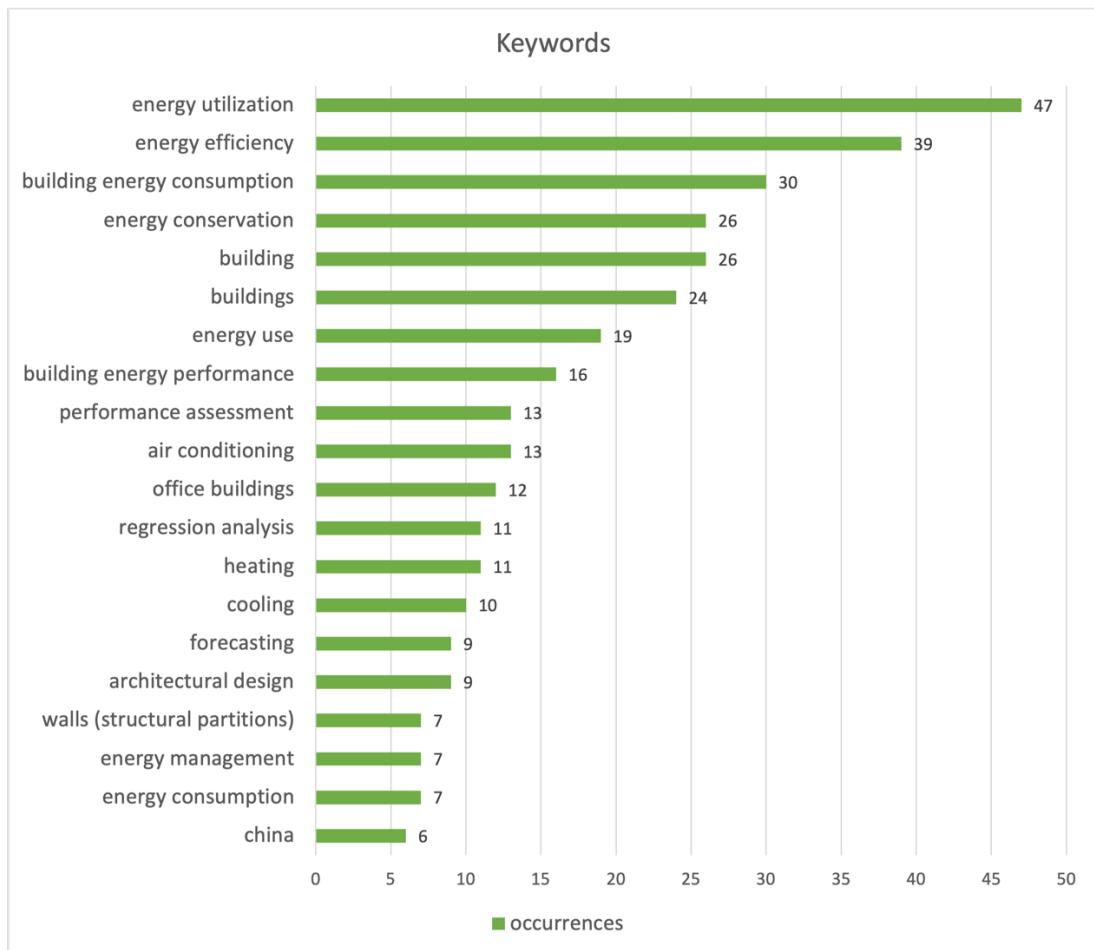


Figure 4.4: Top 20 keyword and number of occurrences

4.4.3 Global Collaboration Analysis

Figure 4.5 shows a global collaboration of the different countries conducting research on the features affecting BEP. Figure 4.6 displays the proportion of publications by the country of publication. China had 22 papers related to this field of research, accounting for 29.73% of all interrelated publications, trailed by the United States with 11 publications (14.87%). However, the average citations values for articles published in the China and United States do not differ significantly, considering the United States produced half as many papers as China. The average citations value of China and United States are 592 and 494 respectively. The total number of citations is arguably the best indicator of quality (Amsterdamska and Leydesdorff, 2005; Leydesdorff et al., 2016), in this regard, China is currently leading in the quality of research published in relation to the features affecting BEP.

These occurrences from China and United States could be due to certain projections for example China is estimated to account for one-fourth of building electricity consumption in buildings in the world by the year 2040 (Capuano, 2019), the United States is projected to beyond 1.3% of total energy yearly (EIA, 2020). Other countries such as the Netherlands and United Kingdom (UK), and South Korea also conducted good quality research. Nevertheless, all countries have a rising quantity of citations which denotes quality contribution to academic literature.

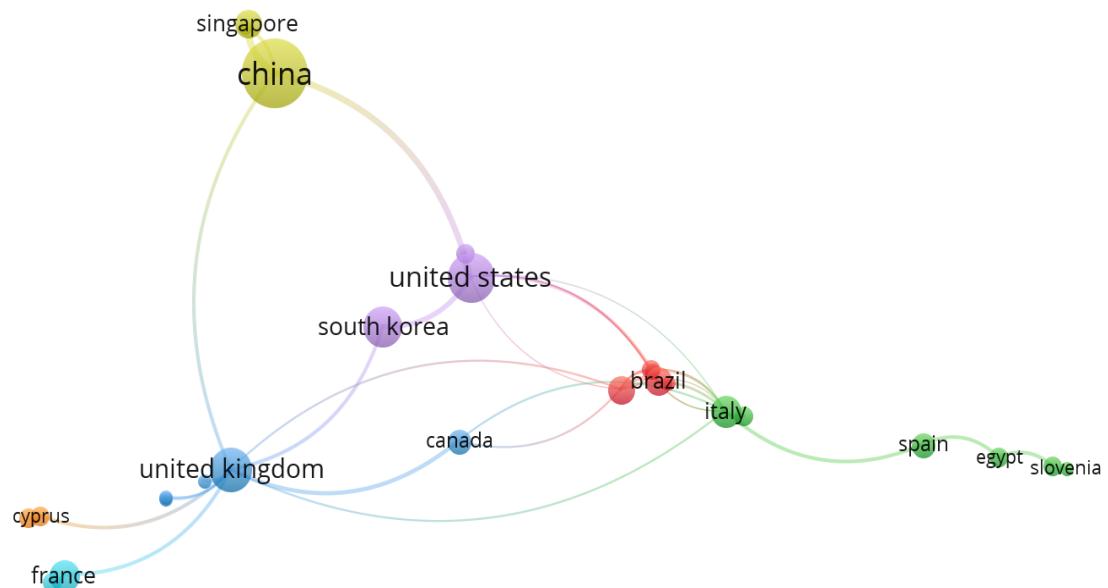


Figure 4.5: Global collaboration network

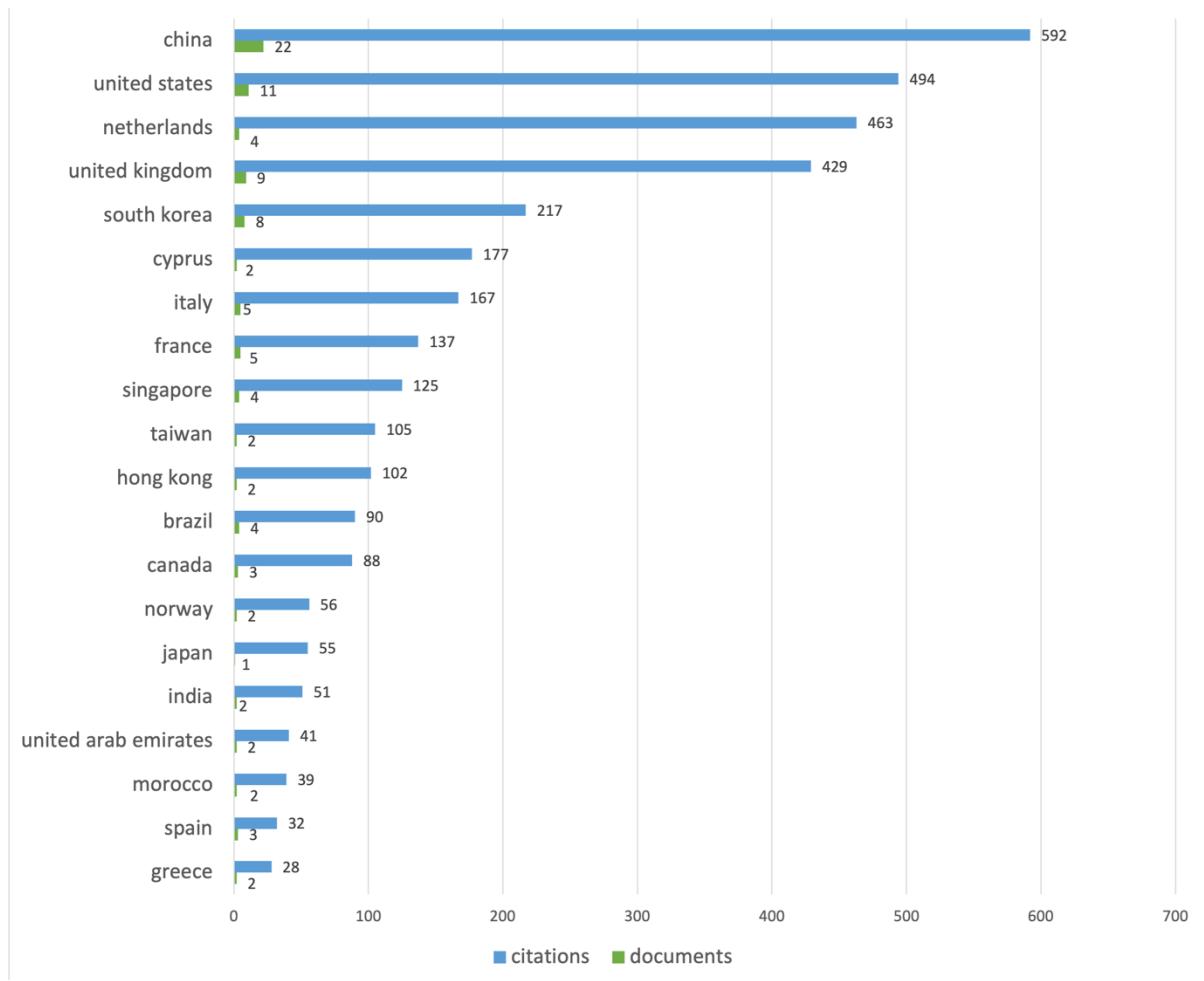


Figure 4.6: Proportion of publications by country.

4.5 The Features

In this research, the influence of various popular and prevalent features of BEP are examined. Numerous studies have explored various features (Fan, 2022; Wang et al., 2020; Yuan et al., 2017), however, seven of the most explored features affecting BEP were reviewed. These features include Building floor area, Windows, Roof, Wall, Weather, Occupancy and Building Orientation. A short description of these features is detailed in Table 4.2 below:

Table 4.2: Description of nine popular and prevalent features of BEP

	Features	Description	References
1	Building floor	Building floor area can defined the overall of footprint range of the buildings inside the buffer zone, multiplied by the relevant quantity of floor levels (Islam et al., 2016). This driver includes the review of building floor and other related	(Ihara et al., 2015; Kim and Suh, 2021; Lee and Lee, 2009; Li et al., 2014; Paukštys et al., 2021)

		building floor features that influence BEP such as number of floors and floor usage.	
2	Windows	Window is a part of a building envelope that opens in the side of a building, which aids the interaction between the exterior and interior environment. (Yapa, 2001) This driver includes the review of windows and other related windows features that influence BEP such as window glazing, insulation among others.	(Qiu et al., 2020; Su et al., 2021; Suh and Chang, 2014; E. Wang, 2017; Yun and Steemers, 2011)
3	Roof	Building roof is known as the body of the building, which is continually influenced by atmospheric agents during the day. The importance of roof has amplified owing to its large area and energy waste from the roof (Beykzade and Beykzade, 2019).	(Aboelata, 2021; Gros et al., 2014; Košir et al., 2018; Yuan et al., 2017)
4	Wall	The wall of a building is the central interface between the building interior and exterior which is also the main channel for the heat exchange between the building interior and exterior. Thus, the reduction of wall energy use is one of the key means to decrease the energy consumption of the traditional building(Z. Y. Li et al., 2018). This driver includes the review of wall and other related wall features that influence BEP such as walls solar absorptivity, wall insulation and thickness among others.	(Gros et al., 2014; Suh and Chang, 2014; E. Wang, 2017; Yun and Steemers, 2011)
5	Weather	Weather is defined as the condition of the atmosphere, to the degree that it is hot or cold, wet or dry and so on. Generally, Weather represents the day-to-day temperature and precipitation activity (Krishnamurthi et al., 2015).	(Fan, 2022; Mafimisebi et al., 2018; Ocampo Batlle et al., 2020; Wang et al., 2020)
6	Occupancy	Building occupancy is the foundation for operations and management of a building. With the growing prerequisite for building energy conservation, occupancy forecasting has become an essential input for simulations.(Jin et al., 2021)	(Lee and Lee, 2009; Li et al., 2014; Wang et al., 2020; Wu et al., 2020)

7	Building Orientation	Building orientation is the alignment of a building in relation to seasonal difference in the sun path as well as dominant wind pattern. The effect of building orientation varies from thermal comfort to ventilation, energy usage. (Ifeoma and Akande, 2021)	(M. Alwetaishi and Benjeddou, 2021; Ghosh and Neogi, 2018; Ihara et al., 2015; Suh and Chang, 2014)
---	----------------------	---	---

4.6 Result and Discussion

This section reveals the result, findings and discussions of the systemic literature review. The results are displayed in form of tables and statistical illustrations. The features affecting BEP were selected from the systematically reviewed articles, each driver was explored by at least one of the reviewed articles. For comparison, the frequency of application of each driver was visualized in Figure 4.7. Considering a large majority of the review studies did not develop a BEPM, except (Ahmad et al., 2018; Li et al., 2020; Wang et al., 2020). Therefore accuracy based on features utilised was not studied. The plot in Figure 4.7 displays the actual frequency based on the number of application (Green bar), while the second bar (blue bar) calculated based on the consideration of the most used driver as 100% frequency of application. Subsequent discussions were based on the second bar in the interest of simplicity.

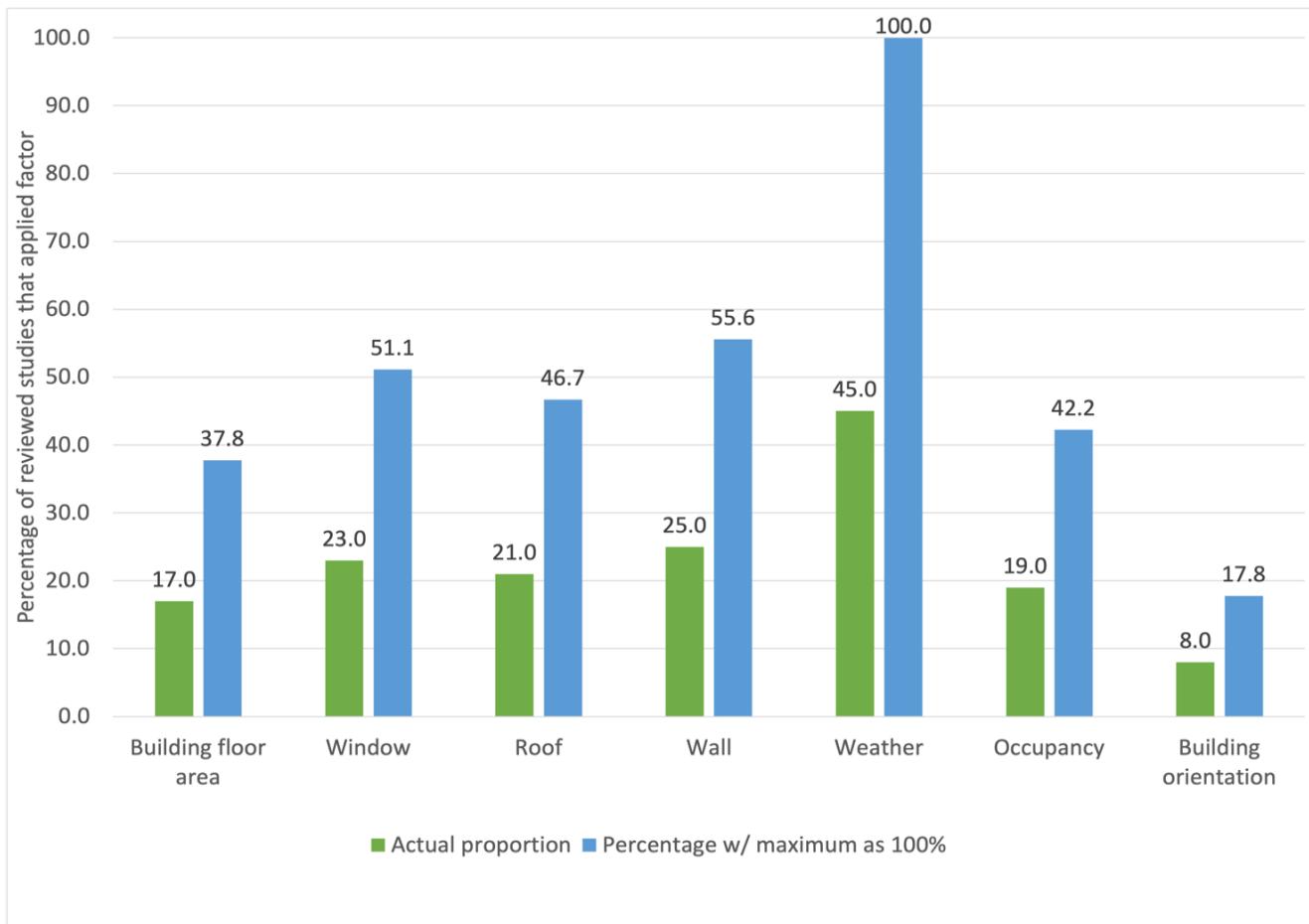


Figure 4.7: Frequency of application of driver in study.

Table 4.3 shows the ranking of the features based on the bar plots in Figure 4.7. According to second bars (blue bars) in the plot, there were only three features out of seven had beyond 50% frequency of application and are measured as the most essential based on the simple evaluation. These features include weather, wall, windows. Of the seven features, it should be noted that only roof driver (46.7%) and occupancy (42.2%) is relatively close to the 50% value. Additionally, it was evident that construction year driver achieved a low frequency rating considering only two studies applied the driver. Thus, it was exempted from the ranking in Table 4.3. The advantages and drawbacks of these features can further corroborate and justify their importance.

Table 4.3: Features and associated ranking

Driver	Above 50%	Ranking
Weather	Yes	1
Wall	Yes	2
Windows	Yes	3
Roof	No	4
Occupancy	No	5
Building floor area	No	6
Building Orientation	No	7

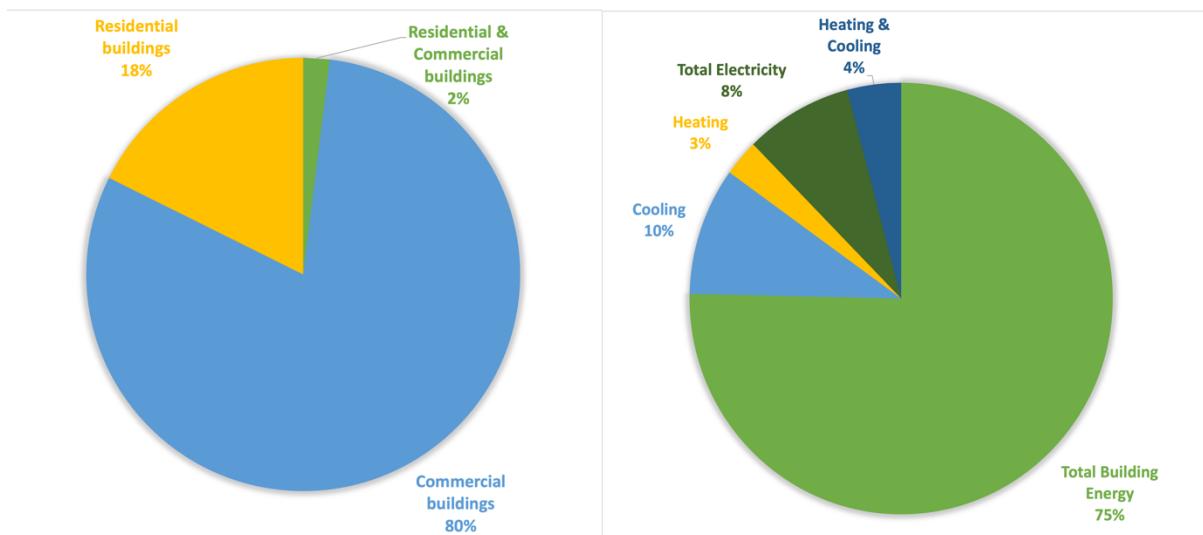


Figure 4.8: Percentage of reviewed articles based on (a) building types, (b) energy types

Figure 4.8 indicates only 18% of reviewed articles focused on investigating features affecting BEP in residential buildings, 80% of reviewed articles centred on commercial buildings while only 2% explored both residential and commercial buildings. Regarding energy types, a substantial fraction of the reviewed articles focused on analysing Total Building Energy, which is a total of 75% while 10%, 8%, 4% and 3% of reviewed articles focused on the cooling, total electricity, heating and cooling and heating respectively. The extent of the entire reviewed articles is shown in Table 4.4, based on features explore

Table 4.4: Data properties, purpose, and features explored in reviewed studies.

S/N	Reference	Purpose of study	Building type	Energy Type	BFA	Window	Roof	Wall	CY	Weather	Occupancy	Orientation
1	(Lin et al., 2018)	Empirical analysis of features of energy use	Residential/commercial	TBE	✓	✓			✓			
2	(Fan, 2022)	Analysis of the Impact of meteorological year	NCS	TBE						✓		
3	(Ocampo Batlle et al., 2020)	Analysis of features influencing energy consumption	NCS	TE						✓		
4	(Wang et al., 2020)	proposed method for building energy prediction	NCS	TBE						✓	✓	
5	(Wu et al., 2020)	Analysis of the impact of occupancy profiles,	NCS	TBE							✓	
6	(Yuan et al., 2017)	Parameter ranking of its influence on BEP	NCS	TE			✓	✓			✓	
7	(Košir et al., 2018)	Analysis of building energy use and indoor thermal conditions	commercial	TBE			✓	✓				
8	(Mafimisebi et al., 2018)	Analysis of features affecting BEP	commercial	TBE						✓		
9	(Suh and Chang, 2014)	Estimation of the energy efficiency of designs	NCS	Heating		✓		✓				✓
10	(Paukštys et al., 2021)	Determine the change in heat loss of in buildings	NCS	Heating	✓							
11	(Li et al., 2014)	Study the influence of climate on BEP	commercial	TBE	✓					✓	✓	
12	(Park et al., 2020)	Demonstrated the importance of the accurate meteorological data	residential buildings	TBE						✓		
13	(Xu and Zou, 2020)	Identified features affecting BEP	commercial	TBE							✓	
14	(Min et al., 2015)	Investigates the net energy consumption	NCS	H&C						✓		
15	(Aboelata, 2021)	Studied extensive and intensive green roofs.	NCS	Cooling			✓			✓		
16	(Yun and Steemers, 2011)	Investigates the behavioural and physical parameters influencing BEP	residential buildings	Cooling		✓		✓		✓	✓	
17	(Lee and Lee, 2009)	Examines the effect of scale features	commercial	TBE	✓					✓	✓	
18	(Peng et al., 2021)	Studied the features influencing air conditioning energy use	commercial	Cooling						✓		

19	(Karatzas et al., 2020)	Analysed occupant behavior patterns	commercial	TBE						✓		
20	(Meng et al., 2020)	Explored effect of weather features on building energy use	NCS	Natural Gas						✓		
21	(Ren et al., 2021)	Examined window operating behaviour	commercial	TBE		✓				✓		
22	(Su et al., 2021)	Analysis of Photovoltaic (PV) windows	commercial	TBE		✓						
23	(Gros et al., 2014)	Analysed impact of cooling materials on BEP	NCS	TBE			✓	✓				
24	(Qiu et al., 2020)	Analysis of Photovoltaic (PV) windows	commercial	TBE		✓				✓		
25	(Chen et al., 2020)	Demonstrates the roles of social-psychological features which influence BEP	commercial	H&C							✓	
26	(Akbar et al., 2020)	Benchmark daily electricity use of a building	commercial	TE						✓	✓	
27	(Mavromatidis et al., 2014)	Proposed methodology to optimize the daylight potential	residential buildings	TBE						✓		
28	(E. Wang, 2017)	Addressed the research gaps to decompose building energy factor structure	commercial	TBE		✓		✓				
29	(Mocanu et al., 2016)	Investigates two stochastic models	residential	TE		✓						
30	(Ahmad et al., 2018)	Investigates the accuracy and generalisation of deep highway networks (DHN)	NCS	TBE						✓		
31	(Li et al., 2020)	ML models comparison	NCS	TE						✓		
32	(Melo et al., 2014)	Evaluate the capabilities of artificial neural network (ANN)	commercial	TBE			✓	✓				
33	(Kim et al., 2021)	Identifying features of the dynamic energy performance gap	commercial	TBE	✓	✓	✓	✓				
34	(Ihara et al., 2015)	Examine the effects of four fundamental facade properties	commercial	TBE	✓	✓	✓					✓
35	(Ghosh and Neogi, 2018)	Examines the effect of geometrical features	commercial	TBE	✓	✓	✓	✓		✓		✓
36	(L. Y. Zhang et al., 2017)	Examines the effects of building external wall's insulation thickness	commercial	TBE				✓		✓		
37	(Y. Liu et al., 2020b)	Develop a support vector machine (SVM) method to predict energy consumption	commercial	TBE						✓		
38	(Chiradeja and Ngaopitakkul, 2019)	Studied the materials and compositions used in building envelopes	commercial	TBE			✓	✓				

39	(Bijarniya et al., 2020)	Examines the effects of various environmental features on the cooling performance	NCS	Cooling			✓			✓		
40	(Ma et al., 2015)	Calculate embodied and operating energy	commercial	TBE	✓							
41	(S. K. Verma et al., 2022)	Examines building materials and ventilation system	commercial	Cooling				✓				
42	(M. Alwetaishi and Benjeddou, 2021)	Examines the influence of window to wall ratio on energy load	commercial	TBE			✓	✓		✓		✓
43	(Jang et al., 2021)	Analysed outdoor and indoor data collected from buildings	commercial	H&C	✓					✓	✓	
44	(Noh et al., 2017)	Proposed a data cube model	commercial	TBE		✓	✓	✓		✓		
45	(Rusek et al., 2022)	To understand how energy is consumed	NCS	TBE						✓	✓	
46	(L. Y. Zhang et al., 2017)	Examined variables that can be applied during building design	residential	TBE		✓	✓	✓				
47	(Geraldi and Ghisi, 2022)	Propose a method to integrate thermal satisfaction into energy benchmarking	commercial	TBE	✓							✓
48	(Zhu et al., 2021)	Investigated the importance of various environmental features	commercial	TBE	✓	✓	✓	✓		✓		
49	(Tafakkori and Fattahi, 2021)	Proposed window configurations for energy efficiency	NCS	TBE	✓						✓	
50	(Liu et al., 2022)	Proposed a data-driven approach	commercial	TBE	✓	✓						
51	(Sooyoun Cho et al., 2019)	Proposed a data-driven approach	commercial	TE						✓	✓	
52	(Gul and NezamiFar, 2020)	Factor analysis	residential	TBE						✓	✓	
53	(Lu et al., 2013)	Selected features to analyze energy utilization indicators (EUIs)	NCS	TBE	✓	✓		✓	✓	✓	✓	✓
54	(Ahn et al., 2016)	Examined the issues that require a solution for application of BEPS tools	commercial	TBE						✓		
55	(C. Robinson et al., 2017)	Proposed a novel technique for building energy estimating learning models	commercial	TBE						✓		
56	(Florides et al., 2002)	To examine measures to reduce the thermal load.	residential	Cooling		✓	✓	✓				✓
57	(Ciulla et al., 2016)	Extrapolate a set of simple correlations of heating energy demand for office buildings.	commercial	TBE		✓		✓				

58	(H. Chen et al., 2021)	Studied the influence of building envelope parameters on BEP	commercial	TBE		✓	✓	✓			
59	(Li et al., 2014)	Studied the influencing features of thermal behaviour of the roofs	residential buildings	TBE			✓			✓	
60	(Iken et al., 2019)	Investigate outdoor wall layer on the BEP	NCS	TBE			✓	✓		✓	
61	(Laaroussi et al., 2020)	Presented main key issues and features affecting the energy behavior	NCS	TBE		✓			✓	✓	✓
62	(Berardi et al., 2018)	Evaluate how temperature affects thermal conductivity of materials in building components	commercial	TBE				✓			
63	(Yu et al., 2022)	To calculate weight coefficients of each subfactor.	commercial	TBE				✓			
64	(Fernandez-Antolin et al., 2019)	Explores the energy efficiency and optimized architectural design	residential buildings	TBE				✓		✓	✓
65	(Yuan et al., 2020)	Examines main features causing overheating	commercial	TBE	✓	✓				✓	
66	(Z. Y. Li et al., 2018)	Triple-glazed window filled with PCM (TW + PCM) is proposed	NCS	TBE		✓				✓	
67	(Huang et al., 2013)	Investigates the building walls in cooling dominate cities	commercial	Cooling		✓	✓				
68	(William et al., 2021)	A sensitivity analysis is undertaken to assess the key features affecting BEP	commercial	TBE					✓	✓	
69	(Mazzeo and Kontoleon, 2020)	Investigated phase change material, green and cool roofs	NCS	TBE			✓			✓	
70	(Yang et al., 2021)	Investigates the interaction effects of occupant behavior-related features	commercial	TBE					✓	✓	
71	(Hasan and Defer, 2019)	To understand the most important features affecting BEP	NCS	TBE						✓	
72	(Pang and O'Neill, 2018)	Investigates key influencing features of BEP.	commercial	TBE	✓						
73	(Rouleau et al., 2018)	Identify the parameters affecting BEP.	commercial	TBE	✓		✓	✓		✓	
74	(Skeie and Gustavsen, 2021)	Examines weather data	NCS	TBE						✓	

Acronyms: - TBE: Total Building Energy, NCS: Not clearly stated, BEP: Building Energy Performance, TE: Total Electricity, H&C: Heating and Cooling. ✓: factor explored in study, CY: Construction year, BFA: Building floor area

The reviewed articles stated the various impacts of the identified features on BEP. Although some studies corroborated the significant effect of the identified features on BEP, some studies stated some opposing arguments of certain features. Therefore, each driver is furthered discussed based on the stipulated theories in reviewed articles.

4.6.1 Impact of Building Floor

In relation to building energy performance, it is practical to infer that large buildings are liable to use more energy in comparison to smaller buildings. Though, it is worth noting that this is merely a typical pattern as it is still conceivable for a large building to be more energy efficient than a small building, if design and operation is conducted with this target in mind (Li et al., 2014). The pearson correlation conducted by (Li et al., 2014), shows a weak correlation between BEP and floor area with a coefficient value of 0.32, indicating no clear or solid proof that small buildings will use less energy than larger ones. The benchmarking of BEP considers a broad variety of different features, including floor area, climate condition among others(Lee and Lee, 2009). (Ling et al., 2015) found that building floor area constitute a large share in space heating loads of residential buildings. Also, the quantile regression results by (Liu et al., 2022), revealed that most features imposed strong diverse effects on BEP. However, in comparison with other consider features, floor area engendered the highest positive effect on energy consumption.

Another essential driver of BEP is the floor usage, or the manner of action or activity conducted in various building areas or spaces. For example, commercial areas such as restaurants or retail stores, or residential spaces such as luxury or flamboyant dwellings, are key features affecting BEP levels. Intriguingly, whilst various studies have highlighted the influence of floor characteristics such as floor area (Lee and Lee, 2009; Li et al., 2014; Paukštys et al., 2021), they did not unequivocally highlight the significance or importance of the floor usage for analysing BEP. Nevertheless, it is noted that different building types (i.e., residential, commercial, industrial) ultimately have varying BEP levels(EIA, 2020). Thus, it is not unexpected to devise floor usage, which reveals the type of operation of the reviewed buildings, to be an influencing driver of BEP.

The geometrical features of a building such as number of floors, area of floor ratio were measured and it was noted that the five-floor models was more susceptible to heat transmittance

through the roofs (Ihara et al., 2015). Therefore, the number of floors is considered the most vital feature that has substantial influence on energy use (R. Wang et al., 2018). It was also noted by (Premrov et al., 2018), that the number of floors is the most effective driver that influence the annual heating energy demand.

4.6.2 Impact of Window

According to the study of IEA, BEP of office buildings is primarily influenced by four features, specifically weather features, envelope performance, occupant behavior and equipment performance. However, one of the greatest common occupant behaviors in buildings is the window operating behavior (Yoshino et al., 2017). Based on the observation and analysis of 35 buildings, It was found that strong correlation lies between temperatures (outdoor weather temperature and indoor air temperature) and window operation (Ren et al., 2021).

In warm areas of Asia, an investigation into the relationship between energy savings and window properties inferred that, windows with low U-values (e.g., triple glazing windows) decreases energy use and likewise contribute to decreasing the total cooling and heating demands (Ihara et al., 2015). Buildings encompassing windows with lower U-values exhibit better energy efficiency and low U values can be reached by material-based solutions such as multi-pane glazing systems and well-insulated frames which are commercially accessible. However, regardless of level of U-value, smaller window have less impact on decreasing annual energy demand(Ihara et al., 2015).

Despite the impact of window glazing, only a few studies explicitly considered and emphasized the essentiality, potential benefits or drawbacks(Ghosh and Neogi, 2018; Ihara et al., 2015; Laaroussi et al., 2020; Ren et al., 2021). Ghosh and Neogi, (2018) stipulated that although glazed facades are progressively being utilized in contemporary buildings to ease interior daylight accessibility and also to beautify the building architecturally. The increased glazed facades application is engendering greater solar gain on the internal part of the building, which is gradually becoming a major problem in hot climate regions. (Su et al., 2021) and (Qiu et al., 2020) proposed and improved approach of employing building integrated photovoltaic (BIPV) window to deliver better thermal performance. The development of a building integrated photovoltaic (BIPV) windows is proposed to be of great significance because it does not only generate electric power, but it also decreases heating and cooling loads in buildings

simultaneously. One of the benefits of BIPV window is that it has a great solar heat gain control ability (Su et al., 2021). In southwest China, (M. Chen et al., 2019) assessed a photovoltaic (PV) window and deduced that it could achieve an energy saving ratio of 83%, when PV windows was fitted on the south facing façade.

In the conception of green architecture, it is noted that, in terms of performance and energy efficiency windows are measured as the weakest component or feature of the building envelope. In buildings, they are responsible for the greatest quantity of direct solar gain and thermal bridging. Windows are responsible for about 20% of total heat loss dependent on the outdoor climatic conditions and size of glazing Thus, modifying the window-to-wall ratio (WWR) can engender considerable influence on energy compared to modifying the external walls' thickness(M. Alwetaishi and Benjeddou, 2021) Therefore, it is recommended that analysis and configuration of WWR should be conducted at the design phase to enhance the BEP(Chiesa et al., 2019).Additionally, Alwetaishi and Benjeddou, 2021 proffered that WWR should ultimately range between 30% to 35%, except in high altitude mountains where intensity of outdoor temperatures is low. This is estimated to bring about definitive results regarding energy consumption.

Furthermore, windows are essential features and significant energy savings can be generated when extra procedures are implemented(Florides et al., 2002). The study by (B. Chen et al., 2021) demonstrates that the U-value of external window, wall, and roof, including the window-to-wall ratio have positive correlation with BEP during winter, while the external wall and roof solar absorptance have negative correlation with BEP. It was further concluded that the proposed solutions with high potential to decrease the energy consumption is low U-value, high solar absorptance material and small WWR(B. Chen et al., 2021).

4.6.3 Impact of Roof

Building roof is considered as one of the most essential structural components of the building in a hot environment and it is estimated to bring about 19% of energy savings when properly insulated(Florides et al., 2002).Various studies have investigated the impact of roof on BEP and proposed effective solutions to reduce energy consumption in buildings(Aboelata, 2021; Košir et al., 2018; Yuan et al., 2017). Cool roof is proposed and substantiated as an effective solution for decrease in heating load of the building during the winter season (Košir et al., 2018). Some other solutions such as green roof is proffered enhance thermal performance and

decrease cooling energy demand in buildings(Aboelata, 2021) Nonetheless, (Gagliano et al., 2015) investigated further by conducting a numerical comparative analysis of all aforementioned types of roof, namely cool roof, green roof and standard roof. It was deduced that cool and green roof provide greater energy savings potential as well as environmental benefits than exceedingly insulated standard roof.

Although there are various innovative roof technologies such as cool roof, green roof and phase change material have been proposed as effective solutions in relation to energy efficiency. Effective designs of standard or traditional roof has become a prerequisite to restrict the utilization of technically convoluted and expensive technologies(Mazzeo and Kontoleon, 2020). The materials expended in construction of roofs have many diverse thermal properties dependent on composition. Hence, it is up to the architect or building designer to select the configuration suitable for the building envelope of each building, as insulator also has a substantial impact on heat transfer. (Chiradeja and Ngaopitakkul, 2019) For this perspective, the design of thermally efficient building roofs can be considered a fundamental element to offer substantial energy savings and environmental benefits in buildings.

4.6.4 Impact of Wall

In contrast to roofs, the effect of the external walls solar absorptivity has not received as much attention in research(Pisello et al., 2017), considering the relatively high number of studies on the importance external walls for building envelope and its significant contribution to building energy consumption(Gros et al., 2014; Suh and Chang, 2014; Yuan et al., 2017; Yun and Steemers, 2011). This is anticipated because walls have low exposure to the sky, However, results from research shows that solar absorptivity can constitute a substantial influence on total energy consumption in buildings (Košir et al., 2018).

In cold climate region, the application of appropriate quantity of wall thickness or insulation is considered an effective approach for reduction of energy consumed in buildings(L. Y. Zhang et al., 2017). For example, (L. Y. Zhang et al., 2017) showed that intensification of insulation thickness has a substantial effect on the heating energy consumption of the building, though it shows a comparatively minor effect on the cooling energy consumption of the building. In the study by (Suh and Chang, 2014), the increase in wall thickness to 250 mm from 50 mm influenced the reduction in heating energy load. Although the proper insulation of walls often

elicits energy saving, it tends to be costly. Despite this increase in insulation cost, the energy savings influence of wall insulation and thickness has become eminent in certain regions such as Greece(Kolaitis et al., 2013) However, studies like (Ihara et al., 2015) focused on office buildings, stipulates that thickness in walls do not necessarily always guarantee energy savings.

There is significant correlation between wall and location weather as the increase wall thickness in cold cities often have significant effect on energy consumption of the building (Huang et al., 2013; L. Y. Zhang et al., 2017). Research on the evaluation of the significance of wall in relation to energy savings, indicates that energy savings in building defer in different climatic zones(L. Y. Zhang et al., 2017). For example, two cities (Harbin and Guangzhou), Harbin is situated in relatively extreme cold region, where the heating energy use accounts for a sizeable fraction of total energy use while Guangzhou is located in the hot summer and warm winter zone, where the heating energy use is very minor, and the annual energy use primarily comprises of cooling energy use. In Harbin, increase in wall insulation thickness elicited significant increase energy savings in the building. While in Guangzhou, increase in wall insulation thickness exhibited insignificant deviation in energy saving of the building(L. Y. Zhang et al., 2017). Furthermore, more often than not, the configuration of the wall of a building is frequently decided by aesthetic and structural deliberations at the design phase. Hence, the increase in probability of high energy demand if the U-value is not sufficiently taken into consideration(Ihara et al., 2015)

4.6.5 Impact of Weather

Weather data is considered one of the major features to accurately predicting energy consumption in building and evaluation of indoor environment (Košir et al., 2018; Park et al., 2020). It is noted that building energy performance is predicated on several fundamental features that define its energy use, especially outdoor temperature, which has been established as a fundamental driver that influences BEP (Mafimisebi et al., 2018). Various other environmental features are closely connected to energy use in buildings such as radiation, building envelope composition among others. However, temperate is considered one of the most impactful features of air conditioning energy load in building(Peng et al., 2021). High temperatures in hot climates brings about human discomfort leading to higher consumption of air conditioning. Hence resulting in a rise in buildings energy motivated by the upsurge in the air conditioning systems operation(William et al., 2021).

Research by (Rouleau et al., 2018) established that among other weather features such as humidity, precipitation, among others, outdoor temperature has the greatest effect on both heating and cool energy load and that solar radiation did not directly influence energy demand in buildings. Also, it is noted that excluding relative humidity, the weather features were more significant for heating in winter than for cooling in summer(Rouleau et al., 2018). (Meng et al., 2020) investigated the impact of weather features on heating energy consumption of educational and healthcare budlings. It was found that educational buildings appeared more susceptible to weather features than the healthcare buildings.

Although weather features are considered very essential driver that strongly affect energy performance(Ihara et al., 2015; Park et al., 2020), it is conclude that even different building of research were situated in the same location and same climate, they would demonstrate significantly distinctive energy performance, which dictates that features other than weather are driving the difference in energy performance. (Li et al., 2014)

4.6.6 Impact of Building Orientation

Not many studies consider building orientation, (Suh and Chang, 2014) argues that building orientation is one of the most influential driver, changing the orientation would have more impact on building performance than increasing or decreasing concrete wall thickness. The selection of architectural features, form and orientation of a building are important decisions made at the design stage of development and can significantly reduce or increase BEP. Thus, it remains paramount to avail designers with frameworks to support decision making for the curation of energy efficient designs (Suh and Chang, 2014) For example, a veranda with orientation positioned northwest and northeast use a larger heating energy load than one facing north owing to solar radiation penetration, which penetrates the north side of the veranda (Suh and Chang, 2014).

The ratio of building heat loss and gain is closely connected with the exposure of the surface area of a building(Florides et al., 2002). Hence, building orientation is considered an imperative driver to be taken into consideration during design stage of buildings and retrofitting projects of a building (Hasan and Defer, 2019; Huang et al., 2013). (Florides et al., 2002) stipulated that the most suitable point of a symmetrical house is directed at four cardinal points and for a stretched house to position the long side facing south. (H. Zhang et al., 2017) also stated that

the correlation between the shape driver and BEP of residential buildings results is often ambiguous due to the thermal action of solar radiation.

4.6.7 Impact of Occupancy

The attention to occupant behavior originated in 1978 in a study by (Socolow, 1978), which examined the impact of residents behavior on space heating energy load. In the last decade, the attention to human behavior has grown in a striking way. Specifically, 2019 was the year with more publications in academic literature. (Laaroussi et al., 2020). (Yang et al., 2021) conclude that the effects of occupant behavior on BEP cannot be ignored, as “Buildings do not consume energy, but humans do”. It is stated that to understand how energy is consumed in building, knowledge of human activity and space occupancy is a prerequisite (Rusek et al., 2022).

It is a concept that if occupants utilize air conditioning systems uneconomically during the summer, there is high probability they will do the same during winter. Likewise, occupants that are used to often opening windows during the summer are inclined to uphold the same pattern in winter. (Rusek et al., 2022). (Yang et al., 2021) investigated the effect of occupant behaviour on energy use in office buildings and it was concluded convoluted relationships lies between occupant behaviors, buildings, climate conditions and equipment. Also, occupancy in residential buildings is an essential feature as there is a direct link between energy use of a building and the occupancy patterns (Rouleau et al., 2018)

4.6 Chapter Summary

The research implemented a systematic literature review research method to highlight the most pertinent factors influencing BEP. It begins with an overview, followed by a detailed elucidation of the research methodology employed. Data collection methods are discussed, including bibliometric analysis techniques such as publication trends analysis, keywords co-occurrence analysis, and global collaboration analysis. The focus of this chapter is on identifying and analysing the various features affecting building energy performance. Results and discussions are provided for each driver, including the impact of building floor, window among others. Results exhibited that three factors namely weather, wall and windows are the most important factors. Through this analysis, insights are extrapolated into the multifaceted factors influencing building energy consumption and efficiency.

CHAPTER FIVE

5.0 RESEARCH METHODOLOGY

5.1 Chapter Overview

This chapter presents the research design and methodology employed in the investigation of the research problem and in the fulfilment of the objectives of this research enumerated in Chapter 1. The purpose of the first section of this chapter is to elucidate the relationship between the ontological, epistemological, and methodological approaches within the research and convey the justification on how they collectively informed the choice of overall quantitative research strategy (5.3), the choice of empirical sample and data collection (5.5).

5.2 Research Philosophy and Approach

The research paradigm or philosophy is an imperative and vital part of research, as it underpins the beliefs and assumptions around the nature of knowledge, which would essentially guide the strategy and method of research (Saunders et al., 2009). Easterby-Smith et al., (2008) stipulated that determining and understanding the research philosophy is measured as the starting point and vital to the research design. This philosophy or paradigm of research will be influenced by practical reflections and the key influence is liable to be a researcher's perspective of the connection between knowledge and method of research. Bryman, (2003, p.4) defined a paradigm as "a constellation of beliefs that influence what should be studied by a scientist, the manner in which the research is conducted and how the result is interpreted". Before further elucidation of the research paradigm, it is necessary to highlight and understand the assumptions utilised in the research philosophy. According to Burrell and Morgan, (1979), whether consciously aware or not, a number of categories of assumptions will be made at every stage of research, which inherently impacts how the research is carried out. These assumptions fall into two key categories namely ontological assumptions and epistemological assumptions (Saunders et al., 2009). Each of these assumptions includes essential distinctions which could influence the way the research processes are understood. Different paradigms contain inherently differing ontological and epistemological views, and so make different assumptions about reality and knowledge. Traditionally, social science research has been approached from two contrasting philosophical traditions: positivism, which is usually referred to as the

quantitative or objectivist approach, and social constructionism, also known as the qualitative, subjectivist or interpretivist tradition (Collis and Hussey, 2003; Easterby-Smith et al., 2008). The next section will compare the different ontological and epistemological views to shed light on the rationale for choosing the positivism approach and the quantitative methodology as the basis of this research.

5.2.1 Positivism and Interpretivism

Positivism is a philosophical stance that is commonly associated with natural science and the study of observable social realities like organizations and managers (Saunders et al., 2009). The goal of this approach is to produce law-like generalizations through the collection of pure, uninfluenced data and facts. The positivist approach is characterized by its use of objective observation, prediction, and testing of causal relationships (Maggs-Rapport, 2001). This paradigm is primarily associated with quantitative research methods and assumes that reality exists independently of humans. Whereas interpretivism mainly relates to qualitative and subjective methods of research (Collis and Hussey, 2009)

While proponents of positivism claim that it is a value-free and objective approach, (Johnson and Onwuegbuzie, 2004) argue that it is subjective, as the researcher must make decisions about what to study, how to collect data, and how to interpret the results. This criticism led to the emergence of post-positivism, which seeks to address the limitations of the positivist approach by recognizing the possibility of the researcher's own beliefs and values affecting the observations(Grix, 2004). Post-positivism recognizes that reality exists independently of the observer, as well as the possibility that the beliefs and values of a researcher could influence the observation.

Interpretivism is a philosophical stance that highlights the subjective and interpretive nature. Hence, a researcher cannot be independent of the research (Saunders et al., 2009). Collis and Hussey, (2009) stipulates that the interpretivism paradigm is considered a subjective, qualitative, and humanist method. Interpretivism is considered a response to the overly dominant effect of positivism. The idea that a single, verifiable reality exists outside of the human mind is dismissed by interpretivism. However, this paradigm has received many criticisms, one of which is being too “soft” to adequately yield theories that could be generalisable to larger populations. Also, for lacking objectivity, based on the researcher’s involvement (Grix, 2004).

5.2.2 Ontological and Epistemological assumptions

Ontology: In social science, the term ontology refers to a researcher's view of the nature of reality, that is, the phenomena under investigation (Saunders et al., 2009). Traditionally, the ontological assumption has two distinct viewpoints: objectivism and subjectivism. Objectivism holds the belief that social entities exist independently, outside of the perceptions and interpretations of the social actors or stakeholders involved. Researchers with a positivist stance tend to view reality as an objective method. On the other hand, subjectivism views social phenomena as being shaped by the perceptions and experiences of social actors (Saunders et al., 2009). Researchers with an interpretive stance are likely to perceive reality as being subjectively and socially constructed. Consequently, different individuals will have varying perceptions of the same situation, which can influence a researcher's decisions and the nature of their interactions with others.

The primary objective of this research is to develop a statistical and AI/ML system for energy assessment at the design stage of residential buildings. This research employs energy data which varies in quality and quantity, independent of human perception. Based on the philosophy that the scope of this work exists independently of human perception(Saunders et al., 2009), the research identifies the key features and tools necessary for model development using objective methods, avoiding dependence on subjective reflection, or intuition.

Epistemology: Epistemological assumption is concerned with the perspective of the researcher on what is considered acceptable knowledge within their field (Cohen et al., 2007; Saunders et al., 2009). It deals with the study of knowledge and what a researcher considers valid and acceptable. Generally, it demonstrates the researcher's view of the world and their relationship with reality and knowledge.

5.3 Research Method

Further to the choice of the research paradigm, the researcher can select the methodology that reflects the philosophical assumptions of the selected research paradigms (Collis and Hussey, 2009). The two often utilised approaches of research methodology are qualitative and quantitative approaches. The traditional understanding of these approaches is that; the

quantitative approach is the analysis of text data while qualitative data is the analysis of numerical data (Easterby-Smith et al., 2001). Qualitative and quantitative approaches vary in different ways (i.e., how data is collected, the nature of the data, the method of analysing the data and interpreting the results) (Haas, 2002). The nature of qualitative research is constructed as looking through a wide lens to discover patterns of correlation between an earlier unspecified set of concepts, while quantitative research looks through a narrow lens at a specified set of variables (Brannen and Coram, 1992).

Qualitative research involves the collection and analysis of non-numerical data, such as text, audio, and video, to understand concepts, opinions, or experiences and it is rooted in interpretivism (Grix, 2004). Qualitative methods often include interviews, and content analysis, aiming to provide comprehensive insights and a deeper understanding of the context in which the behaviour occurs. While qualitative research offers valuable insights and is indispensable in many fields for understanding participant perspectives and generating detailed data, it does not meet the requirements of this research. This research requires precise, measurable, and objective data that can be utilized to train and validate statistical and machine learning algorithms. Qualitative data, being inherently subjective and harder to quantify, would not provide the structured, numerical input necessary for these predictive models. Moreover, this research seeks to leverage energy data and building metadata (numerical) for the prediction of potential energy consumption at the design stage. Therefore, the quantitative method is adopted as the best fit for the research objectives.

The qualitative and quantitative methods each have their advantages and disadvantages, and the choice of each method is predicated on the research question and data type collected. Some of these advantages and disadvantages are highlighted in Table 5.1 below.

Table 5.1: Qualitative vs Quantitative method (Casebeer and Verhoef, 1997; Easterby-Smith et al., 2001)

	Qualitative	Quantitative
Definition	The non-numerical analysis and understanding of observations, with the aim of discovering patterns of	The numerical manipulation of observations with the aim of describing the

	corelation or inter-relationship	phenomena that the observations reflect.
Type of reasoning	<ul style="list-style-type: none"> • Induction • Subjective 	<ul style="list-style-type: none"> • Deduction • Objective
Type of philosophy and approach	<ul style="list-style-type: none"> • Epistemology • Interpretivism 	<ul style="list-style-type: none"> • Ontology • Positivism
Advantages	<ol style="list-style-type: none"> 1. New theories can be generated. 2. Complex questions can be examined which could be impossible with quantitative. 3. Explore new areas of research 	<ol style="list-style-type: none"> 1. It can be easily generalisable. 2. It is an objective method. 3. It utilizes variables which are measurable. 4. It is replicable and can handle large sample size
Disadvantages	<ol style="list-style-type: none"> 1. It is less easily generalisable. 2. It is subjective and this leads le to procedural issues. 3. The bias of the researcher is often unavoidable 	<ol style="list-style-type: none"> 1. It is limited to rigidly definable variables. 2. It is less helpful in theory generation

Being grounded on positivism philosophy and deductive approach, this research adopts the quantitative method. This will help accomplish the aim as statistical and ML algorithms utilize quantitative data (numerical values) for prediction. Also, it is considered the best because it focuses on specific behaviours that can be quantified and does not manipulate variables (Cozby and Bates, 2012). Qualitative research was not chosen because it is the analysis of qualitative data (text data).

5.4 Research Approach

Generally, the validity of the choice of a research approach for a study is predicated on the goal of the study and its research questions (Saunders et al., 2009; Yin, 2003). As discussed in section 2.1, there are different paradigms for conducting research and each of them offers varying perspectives. Based on the research questions (i.e., RQ1 What are the most common features that influence energy consumption in buildings? among others) and goal of this study (i.e., development of a reverse engineered system for energy assessment at the design stage of residential buildings), the positivism paradigm is adopted in this research. This paradigm was selected over interpretivism, realism and pragmatism due to several reasons.

This research requires an objective approach and Positivism is a philosophical stance that views reality objectively and assumes that reality exists independently of human perception (Saunders et al., 2009). Also, the positivist approach emphasized the use of quantitative methods such as statistical analysis to test hypotheses Collis and Hussey, (2009). In the context of this research, Positivism promises unequivocal and accurate knowledge (Saunders et al., 2009). This is particularly essential in the building energy use prediction domain, where accuracy and reliability of predictions are imperative for energy management. Another reason for the choice of the positivism approach is the delivery structured and repeatable methodology for conducting research. Additionally, positivism was selected over post positivism in this research. This is because positivisms rely on empirical observation and analysis to generate reliable knowledge, while post-positivism acknowledges the subjective nature of research which could lead to unreliable results.

More explicitly, other paradigms were not adopted for several reasons. Firstly, interpretivism is a philosophical stance that emphasises a qualitative and subjective method of research (Saunders et al., 2009). It aligns with the assumption that reality is subjective and constructed through human interpretation). This would be more applicable to studies focused on human perception, for instance, the study of experiences of women in the workplace by means of in-depth interviews. In the context of building energy use prediction, the interpretivist approach lacks the objectivity and accuracy required for a predictive model. On the other hand, the realism paradigm highlights an objective approach, however, it emphasizes the comprehension of reality is shaped by human perception and experiences(Saunders et al., 2009).

Finally, the pragmatism paradigm accentuates the importance of delivering a practical solutions that can inform future practice (Saunders et al., 2009). While this approach could be construed

as relevant in this field, Pragmatism highlights the stance of the mixture of qualitative and quantitative within one study. The positivist method provides a more suitable framework for the goal and research questions, as it focuses more on empirical data and scientific methods, which is the level of accuracy required in this research.

Bayona-Oré et al., (2021) employed the positivism method for the development of a price prediction model for agricultural products because it is a quantitative and longitudinal research. Similarly, Mangula, (2019) adopted the positivism approach using the deductive approach for energy access modelling in rural areas of Tanzania. Consequently, this research adopts the positivism paradigm to develop a machine learning model for predicting potential energy consumption at the early design stage of buildings.

5.5 Data Collection

The quest to collect the appropriate data presented several challenges. Initially, data was obtained from various sources with different levels of detail and completeness. For instance, data was collected from LSDPC Nigeria, which involved connecting with a resource and the facility manager, yielding a year-long data set for a single unit. Subsequent efforts included downloading energy data for U.S. buildings from the Kaggle Repository and acquiring detailed half-hourly energy data for 11 university buildings from the University of Hertfordshire (UoH) after contacting the energy manager. Despite receiving the energy data, additional building metadata was sought, resulting in limited success. Further challenges were encountered with Ikeja Electric distribution company (IKEDC), which provided a year-long data of dedicated feeders without building metadata. Subsequently, data was downloaded from Datamill North for over 100 buildings in Leeds, UK, followed by extensive communication with leeds.gov.uk and other entities for building metadata, with mixed responses. Similar efforts involved receiving incomplete data from the Irish Social Science Data Archive (ISSDA).

Additional steps included downloading data from the Dataverse repository for British Columbia buildings, attempting to access domestic energy data from the Department for Business, Energy and Industrial Strategy, and engaging with the European Council for an Energy-Efficient Economy, which did not have the required data. Finally, a student account was created to access comprehensive energy performance building data from the Ministry of Housing, Communities, and Local Government, which included necessary building metadata.

This extensive process highlights the complexity and effort employed to gather and compile accurate and comprehensive data for building energy consumption analysis.

Further to the systematic literature review, the relevant data was collected based on the features identified. This research utilized three types of datasets for the development of an energy prediction model for the design stage of buildings. This dataset includes building metadata, meteorological and energy datasets. These datasets contain several features considered pertinent in energy performance prediction which have been employed in various studies (Feng and Zhang, 2020; Olu-Ajayi and Alaka, 2021; Jinsong Wang et al., 2021). These datasets were collected for many residential buildings in the United Kingdom (UK). This research uses only residential building data as they for the majority of building stock in the UK.

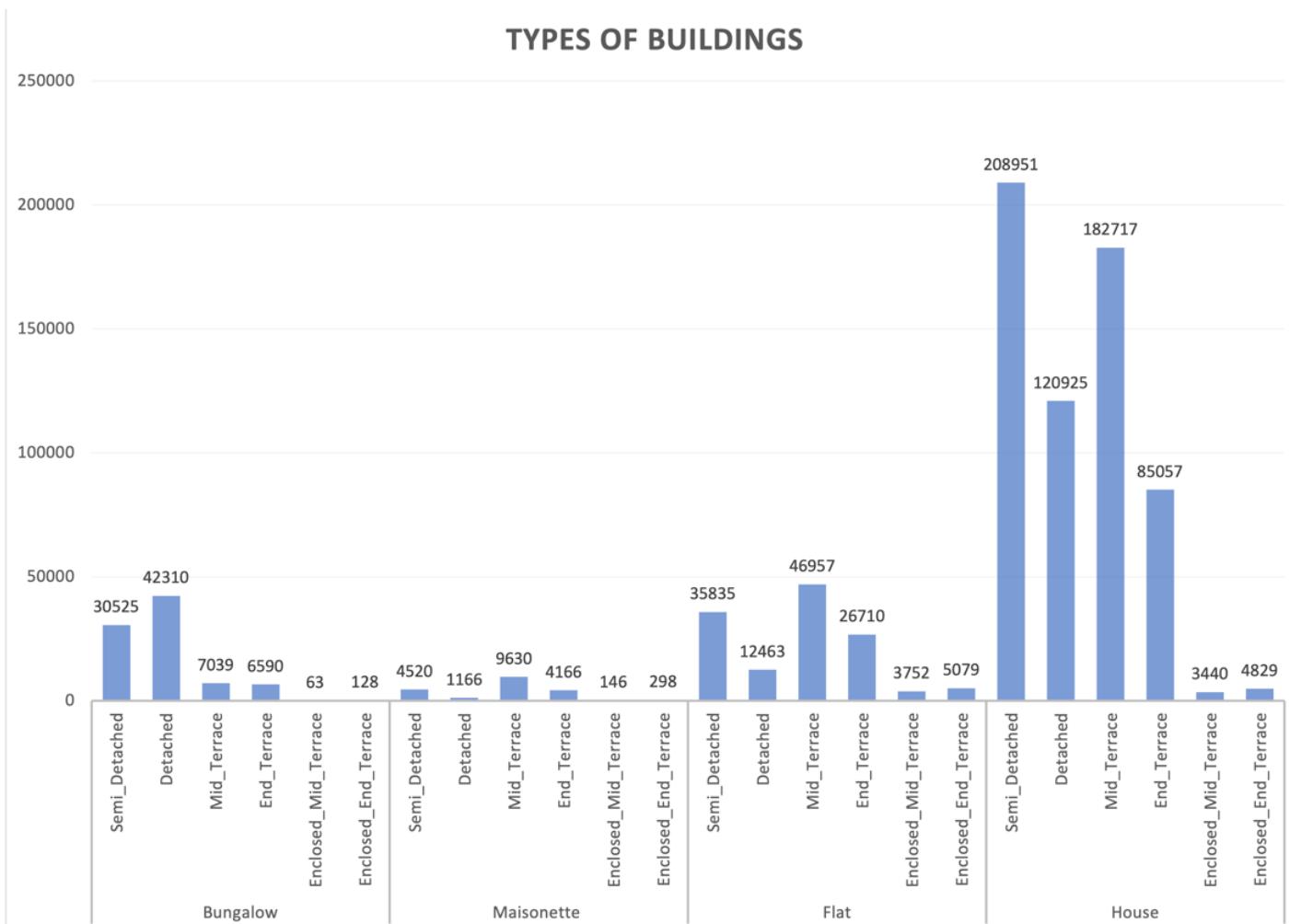


Figure 5.1: Graphical representation of the types of buildings utilized.

The building-related dataset contains the metadata and energy consumption data of 843,296 different types of residential buildings in the UK as shown in Figure 5.1 below.

5.5.1 Building Metadata

The building metadata for 843,296 residential buildings located within 47 area postcodes in the UK were collected from the Ministry of Housing Communities and Local Government (MHCLG) repository. The data consist of 843,296 residential buildings from forty-seven (47) different postcode areas (Blackburn, Blackpool, Darlington, Yorkshire, Halton, Hartlepool, Hull, Middlesbrough, Redcar Cleveland, Warrington, among others) for a clear and less ambiguous comparison. The building metadata comprised of only parameters that can be detected and modified during the design stage. This includes floor level, roof description, wall description and Number of Habitable Rooms among others as shown in Table 5.2 below. These were considered as the independent variables while the energy data was considered as the dependent variable. The input variables such as wall and window type or description are considered important variables as they have an effect on the energy consumption of buildings (Marino et al., 2017; Marwan, 2020; Tahmasebi et al., 2011).

Table 5.2: Building and Weather-related variables collected.

S/N	Variable	Description	Internal/ External	Type	Label
1	Temperature [°c] (Annual Average)	Weather temperature is the degree of hotness or coldness of a particular area.	Meteorological Data	Continuous	Independent
2	Wind speed [km/h] (Annual Average)	This is the rate at which air is moving in a particular area.			
3	Pressure [Hg] (Annual Average)	Air pressure is the force exerted onto a surface by the weight of the air.			
4	Precipitation	The amount of rain, snow, hail, etc., that has fallen at a given location within a set period			
5	Postcode	The postcode of the property	Building Data	Categorical	
6	Total Floor Area[m ²]	This is the total of all enclosed spaces measured to the internal face of the external walls measured based on the Royal Institute of Chartered Surveyors guidance issued.			
7	Property Type	This describes the type of property (i.e., Flat, House, Maisonette etc).			
8	Glazed Area	This describes the total glazed area of the habitable area. e.g., much less than typical, less than typical, normal etc.			
9	Built Form	The building type of the Property e.g. Detached, Semi-Detached, Terrace etc. This combined with the Property Type provides a structured description of the property.		Discrete	
10	Extension Count	The describes the number of extensions.			
11	Walls Description	This describes the wall type (e.g., Cavity wall, Solid brick, Park home wall etc).			

12	Floor Description	This describes the floor type (e.g., Solid, Suspended etc).			
13	Floor Height	The average height of the storey in metres.		Discrete	
14	Windows Description	This describes the window type (e.g., Cavity wall, Solid brick, Park home wall etc).			
15	Windows Energy Efficiency	Energy efficient window glazing covers both double and triple glazing. These are windows with two or more glass panes in a sealed unit. Also, the energy efficiency of a building can be improved by the installation of secondary glazing. Therefore, the window energy efficiency is rated from very good to very poor based on the type of glazing utilized.			
16	Windows Environmental Efficiency	This is concerned with the rating of the quality of window in terms of environmental friendliness. The most important aspect of any window is the glass, a relatively environmentally harmless material made from sand. This is rated from very good to very poor.			
17	Mainheat Description	This describes the mainheat type (e.g., Electric storage heaters, Boiler and radiators, mains gas).		Categorical	
18	Mainheat Energy Efficiency	Good-quality energy efficient mainheat reduces air inflation and blocks the flow of heat. This is rated from very good to very poor.			
19	Mainheat Environmental Efficiency	This is concerned with the rating of the quality of the mainheat in terms of environmental friendliness. This is rated from very good to very poor.			
20	Main Fuel	The type of fuel used to power the central heating e.g. Gas, Electricity			
21	Walls Energy Efficiency	Good-quality energy efficient wall reduces air inflation and blocks the flow of heat. An energy-efficient wall also controls water and vapor intrusion and protect against thermal radiation. This is rated from very good to very poor.			
22	Walls Environmental Efficiency	Through shading, walls can lower temperatures in summer and reduce energy cost. Therefore, the window energy efficiency is rated from very good to very poor based on the type of glazing.			
23	Roof Description	This describes the roof type (e.g., 50m, 100m loft insulation).			
24	Roof Energy Efficiency	Roofs are good ways of conserving energy. This is rated from very good to very poor.			
25	Roof Environmental Efficiency	This describes the environmental efficiency of the type of roof selected. Also, this is rated from very good to very poor.			
26	Lighting Description	This describes the lighting type			
27	Lighting Environmental Efficiency	This describes the environmental efficiency of the type of lighting selected. Also, this is rated from very good to very poor.			

28	Lighting Energy Efficiency	Energy-efficient lighting helps reduce the energy consumption and carbon dioxide emissions, all without decreasing the quality of light in our homes.			
29	Number of Heated Rooms	The number of heated rooms in the property.	Discrete	Categorical	Dependent
30	Number of Habitable Rooms	Number of Habitable rooms include any living room, sitting room, dining room, bedroom, study and similar.			
31	Energy Rating	This is the building's annual energy rating	Building Energy Data	Categorical	Dependent

5.5.2 Meteorological Data

The meteorological dataset was collected from the Meteostat repository, and it contains weather features such as temperature, wind speed and pressure as shown in Table 5.2 above. Figure 5.2 shows the monthly weather temperature data. According to Ding and Liu, 2020, one of the major variables for building energy prediction is meteorological data (Ding and Liu, 2020). This data was collected for 47 area postcodes of the residential buildings and the granularity of meteorological data collected was daily average from 1st January 2020 till 31st December 2021. This duration was selected for weather data because the energy value of related buildings was assessed between the years 2020 to 2021.

The building and meteorological variables are enumerated in Table 5.2 above. The building data is categorized as internal while the meteorological data is categorized as external.

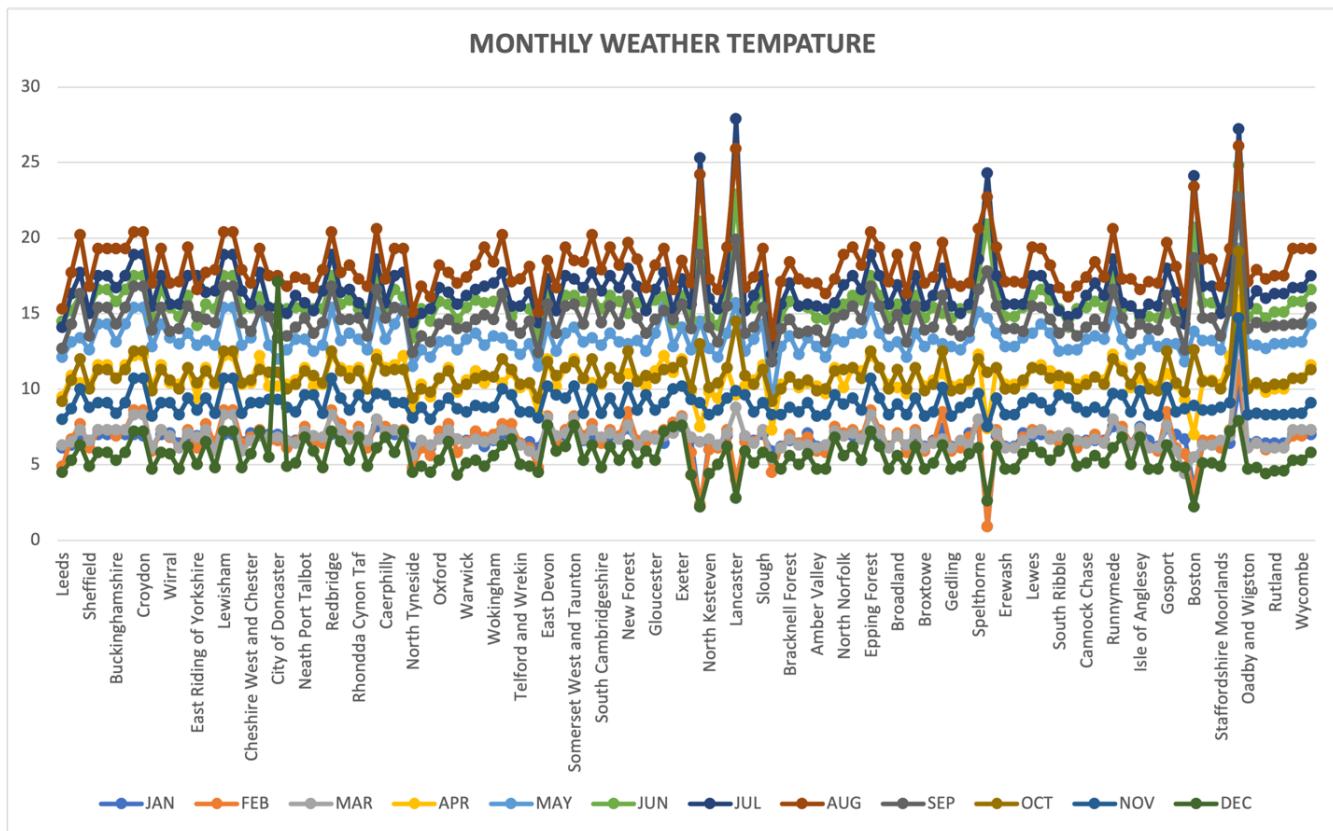


Figure 5.2: Monthly weather temperature.

5.5.3 Energy Data

The energy dataset was also collected from the Ministry of Housing Communities and Local Government (MHCLG) repository, which includes the annual energy consumption of each building for the years 2020 and 2021. Both the energy consumption values, and energy rating were included in this dataset. In the quest to reduce the high proportion of energy in buildings, the European Union (EU) applied a systematic framework for understanding energy performance which has since prompted member states to generate certification systems for rating building energy performance (European Parliament, 2002). Hence, the UK employs a standard scale rating system to notify building owners of current energy ratings, energy cost and effectual recommendations to improve energy efficiency (Curtis et al., 2014).

Energy Efficiency Rating		Current	Potential
Very energy efficient - lower running costs			
(92-100)	A		
(81-91)	B		
(69-80)	C		73
(55-68)	D		
(39-54)	E		
(21-38)	F	37	
(1-20)	G		
Not energy efficient - higher running costs			

Figure 5.3: Energy efficiency rating (gov.uk)

The building data and the meteorological data were employed as the independent variables while the building energy consumption or rating was utilized as the dependent variable.

The building energy rating adopts the UK standard rating which is issued based on the energy consumption level of the building. The energy rating in this data includes A, B, C to G, with ‘A’ denoted as the most energy efficient with the lowest running cost while ‘G’ signified the least energy efficient with the highest running cost as represented in Table 5.3 below. The EPC energy efficiency rating was utilized as the target variable in the development of a classification model, while energy consumption values were utilized in the development of a regression model.

Figure 5.4 shows the proportion of building energy efficiency ratings for each building type. Subsequently, to prevent complexities during model development, the data was pre-processed.

Table 5.3: Energy rating and energy performance values

S/N	Energy performance value	Energy efficiency rating	Remarks
1	0-25	A	
2	26-50	B	
3	51-75	C	
4	76-100	D	Fairly energy efficient
5	101-125	E	
6	126-150	F	
7	Over 150	G	

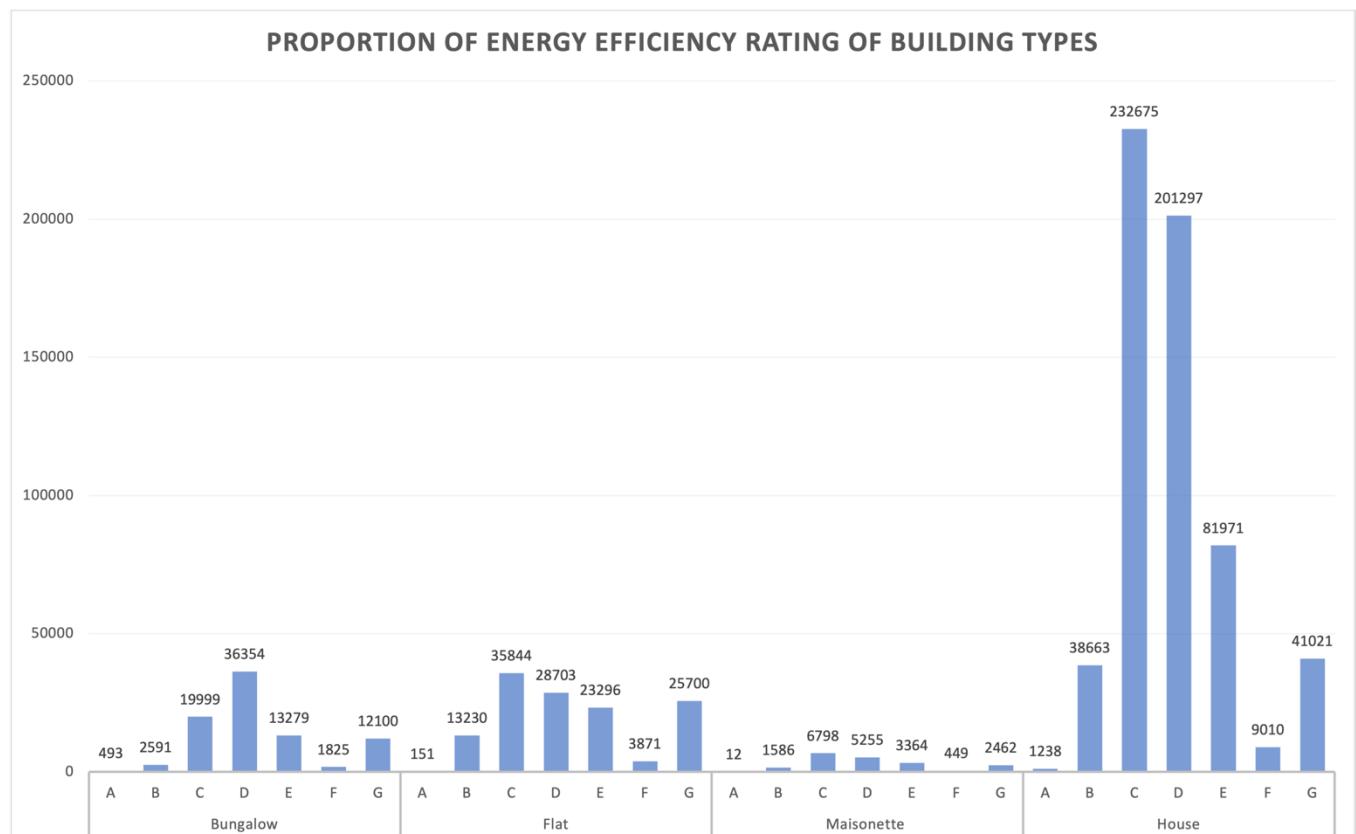


Figure 5.4: Proportion of energy efficiency ratings for each building type.

5.6 Chapter Summary

This chapter provides a comprehensive exploration of the research methodology utilized in this research. This chapter delivers the research philosophy and approach and examines ontological and epistemological assumptions. It also discusses the perspectives of positivism, realism, and interpretivism. Following this theoretical foundation, the positivism paradigm was adopted in this research. Subsequently, A detailed discussion on the data collection method, and methodologies utilized to gather relevant data. Furthermore, the chapter highlights the importance of data pre-processing in ensuring the quality and integrity of the collected data

CHAPTER SIX

6.0 FEATURE SELECTION EVALUATION FOR BUILDING ENERGY CONSUMPTION PREDICTION

6.1 Chapter Overview

This chapter investigates the effectiveness of feature selection centred on building energy consumption prediction. This chapter examines the most relevant features by analysing different feature methods. The identification of the best feature combination has the potential to enhance model performance, reduce computational complexity, and mitigate the risk of overfitting. Additionally, this chapter examines the effect of feature selection methods on both classification and regression prediction model performance and delivers the theoretical and practical implications of such analysis. This chapter will employ various performance measures to evaluate the performance of the classification and regression prediction model. These measures include Root Mean Squared Error (RMSE) Mean Absolute Error (MAE) and Accuracy among others. Additionally, this chapter will comparatively analyse the impact of using the same algorithm for feature selection and prediction against using a different algorithm for feature selection and prediction. Lastly, by conducting various analyses of feature selection and the impact of its methods on model performance, this chapter will address one key hypothesis (**H0**).

6.2 Significance of Feature Selection

One of the most broadly unaddressed issues in energy consumption literature, which affects ML algorithm performance is the feature or variable selection (Hsu, 2015). Identifying the most relevant features strongly affects the accuracy of the predictive model (Alaka et al., 2018; Pirbazari et al., 2019). Several studies often select variables derived from academic literature to develop ML models. However, before model development, the application of the feature selection method using statistical techniques is considered one of the most effective methods for identifying the most important variables or features (Hai-xiang Zhao and Magoulès, 2012a). Feature Selection (FS) aims to improve the performance of ML models by eliminating unimportant and irrelevant noisy features, thus improving the quality of the dataset (Asir et al., 2016). These redundant features are removed to reduce the input dimensionality. It is considered essential to apply feature selection for optimum model performance (Alaka et al., 2018). Researchers have established that the appropriate choice of features or variables is

closely connected to the increase in performance accuracy of a model (Alaka et al., 2018; Balogun et al., 2021; Zhang and Wen, 2019a). Feature selection is recognised as a data pre-processing technique for efficient data preparation (mainly high-dimensional data) in machine learning problems (Asir et al., 2016; J. Li et al., 2017; Maldonado and Weber, 2009). High-dimensional features often incur a high computational cost, while low-dimensional data decreases the probability of overfitting. Feature selection measures the relevance dependency of each feature with the output label (Chandrashekhar and Sahin, 2014). Irrelevant features are features that have no impact on the target function in any way, while redundant features are features that add nothing to the target function (Dash and Liu, 1997). The elimination of irrelevant and redundant variables often reduces the data, leading to enhancement in the classification performance. Additionally, feature selection decreases the computational time as well as the difficulty of training and testing a classification model. Therefore, it engenders more cost-effective predictive models (Effrosynidis and Arampatzis, 2021). Furthermore, selecting the most relevant features simplifies the calculation and reduces the dimensionality (HaiXiang Zhao and Magoulès, 2012).

Generally, a feature is a singular quantifiable property of a process being perceived (Chandrashekhar and Sahin, 2014). In real-world situations, data is often represented using numerous features. However, in most cases of an ML prediction model, only a few of these features correlate to the target output (Kira and Rendell, 1992). The unrelated features that constitute no correlation to target output serve as pure noise, which could lead to bias in prediction, thereby diminishing the classification performance (Kunasekaran and Sugumaran, 2016). To avoid such bias and improve classification performance, feature selection is required to speed up the learning process and enhance the quality of data. In the field of building energy use prediction, features (or input variables) are mostly selected based on domain knowledge, [e.g., (Bagnasco et al., 2015; Bourhnane et al., 2020; Ding and Liu, 2020; Dong et al., 2021a; K. Li et al., 2018)]. However, more recently, the integration of feature selection methods with machine learning for predicting building energy usage has slowly been explored. For example, random forest and Pearson's correlation coefficient were applied to rank a total of 124 features for building energy use prediction (Zhang et al., 2018). Also, (Faisal et al., 2019) utilized recursive feature elimination and mutual information methods to calculate the importance of the input features for predicting electricity consumption. It was concluded that the results produced using the selected features outperformed the results using the original features. Furthermore, it is suggested in the same study that feature selection can aid the reduction of frequently experienced overfitting issues.

Additionally, Zhang and Wen, (2019a) conducted a novel exploration of feature selection methods. Initially, the original feature sets comprised 278 features. However, using domain knowledge 22 features were chosen, subsequently using the Pearson correlation coefficient 29 features were selected. Lastly, multivariate adaptive regression splines were employed for thorough selection which elicited 14 features that led to better model performance. (Paudel et al., 2017) employed feature selection in the prediction of energy consumption and emphasized the benefits of feature selection for an increase in accuracy levels and a decrease in computational times. Previous studies have applied the feature selection method in the development of energy predictive models (M. W. Ahmad et al., 2017; Z. Dong et al., 2021; Zhang & Wen, 2019b). However, the fraction of studies that deliver comprehensive insights on the incorporation of feature selection with machine learning is still limited, notwithstanding the capabilities of feature selection.

For instance, despite the noted importance of feature selection in machine learning prediction models according to a majority of studies (M. W. Ahmad et al., 2017; Z. Dong et al., 2021; Zhang & Wen, 2019b). It has been argued by a few studies that carried out experimental research that feature selection can have negative impacts on ML model performance (Balogun et al., 2021; Kapetanakis et al., 2017). A critical review of most of these studies shows that different feature selection methods were employed, and it can be argued that while feature selection does impact prediction, this impact could be negative or positive depending on the algorithm utilized.

There are three key feature selections established for feature selection namely filter, wrapper and embedded methods (Maldonado and Weber, 2009; Zhang and Wen, 2019b). Filter methods select features centred on performance measures without consideration of the type of modelling algorithm utilized (Jović et al., 2015). Few studies have applied filter techniques to select relevant features in building energy use prediction. For instance, (Kapetanakis et al., 2017)) used a linear correlation feature selection technique in the development of a thermal load prediction model. It was concluded that the accuracy remained on the same level as the FS application. Kusiak et al (2010) also applied the boosting tree feature selection technique in the selection of the relevant variables for predicting building steam load. The relevant input variables included the maximum, minimum and mean of the outdoor air temperature (Kusiak et al., 2010).

Filter methods provide more generality as they are independent of the chosen algorithm. They are fast in execution and computationally inexpensive in comparison to the wrapper and embedded techniques (Asir et al., 2016). The wrapper and embedded methods are considered less efficient than filter methods for high-dimensional data processing (Iqbal et al., 2020). The filter method is implemented based on the common characteristics of features such as distance between classes and statistical dependencies (i.e., each feature is allotted a statistical score). However, this method is blind to any connections between the features. Thus, this method will not recognise features that can be relevant when combined with other features (Bommert et al., 2020). This method removes irrelevant features reducing the feature set dimensionality without losing much model accuracy (Zhang and Wen, 2019b). There are different types of filter methods namely chi-square, boosting tree, linear correlation, and ANOVA, among others.

Filter methods utilize techniques of variable or feature ranking as the customary criteria for feature selection (Aziz et al., 2017). A group of statistical methods are employed to grade each feature or the whole feature set, Contingent on whether multiple features can be assessed simultaneously. Contrasting to filter methods that implement feature selection independently of the development of the prediction model, the wrapper method utilizes an ML algorithm for feature subset evaluation concerning classification error and accuracy.

The major difference between the wrapper and filter is the evaluation criteria. Kohavi and John, 1997 developed the wrapper-based feature selection technique for selecting relevant features (or input variables) from the dataset. The performance of this technique is often evaluated based on classification accuracy using naïve Bayes and decision tree classifiers. However, the wrapper method has difficulties e.g. overfitting, overhead searching and prolonged computational time (Kohavi and John, 1997). Furthermore, the wrapper method utilizes algorithms to assess the produced subsets by using the searching technique, making it more computationally complex. Thus, these techniques are not appropriate for high-dimensional space s(Asir et al., 2016). There are various types of wrapper methods such as recursive feature elimination, and forward feature selection, among others.

Various studies have adopted the wrapper method in the building energy use prediction field. Fan et al., 2014 employed recursive feature elimination to conduct feature selection for eight machine learning algorithms in predicting next-day building energy consumption. After the evaluation of the algorithms, it was observed that Random Forest (RF) and Support Vector Regression (SVR) produced the best result in terms of model performance (Fan et al., 2014a).

Also, Ahmad et al., 2017a utilized random forest filter techniques in the development of Random Forest (RF) and Artificial Neural Networks (ANN) for the building energy use prediction. The filter method produced a better performance in ANN than RF. Likewise, Dong et al (2021) applied the RF feature selection method and employed stacking, ANN and Support Vector Regression (SVR) for forecasting hourly energy use. Stacking emerged as the best among other algorithms (Dong et al., 2021a). Furthermore, Kolter and Ferreira (2011) implemented a forward selection method in predicting energy consumption in a building. The selection of best performance was centred on the Root Mean Squared Error (RMSE) of a predictive model. It was concluded that the RMSE of an energy predictive model can be decreased by utilizing the selected features.

The embedded method assesses the usefulness of features similar to the wrapper method (HaiXiang Zhao and Magoulès, 2012). However, this method implements feature selection during the algorithm's execution; thus, these methods are embedded normally or as an extended functionality in each regression or classification algorithm. The popular embedded methods include some decision tree algorithms such as Embedded Random Forest, Classification and Regression Tree (CART), among others. Furthermore, embedded methods can apply feature selection by the algorithm training process. Thus, due to the abstention of retraining the specific algorithms, it suffers less computational burden. However, it is peculiar to specific algorithms and hence they are not always used (Zhang et al., 2019).

Embedded methods employ the fundamental characteristic of ML algorithms to execute feature selection(Ang et al., 2016). Embedded methods have different methods: during the process of training the model, features with smaller correlation coefficient values are eliminated recursively through the use of a support vector machine. Another method is the application of feature selection as an embedded function during the training process.

6.3 Methodology for Feature Selection

This research analyses the effect of feature selection methods on various ML algorithms in the prediction of energy use in buildings. This analysis will address five key hypotheses (H0-H4) (see Section 1.4.2) related to the effect of openings on building energy consumption, the effectiveness of feature selection in classification and regression prediction, the positive impact of feature selection on model performance, and the comparison between machine learning feature selection and conventional feature selection methods. The schematic diagram of this chapter is displayed in Figure 6.1 below. This section consists of major processes namely:

feature selection, correlation analysis, model training and model testing. This research will develop building energy consumption prediction models using various ML classification and regression algorithms [such as Random Forest (RF) (Carrera et al., 2021; Y.-T. Chen et al., 2019; C. Li et al., 2018; Pham et al., 2020), Support Vector Machine (SVM) (Dong et al., 2005; Jing et al., 2022, p. 202; Y. Liu et al., 2020b; M. Shao et al., 2020; Zhong et al., 2019) etc.], and apply various feature selection methods [such as random forest (Z. Dong et al., 2021; Z. Wang, Wang, Zeng, et al., 2018; Zhang & Wen, 2019b) and chi-square (Bahassine et al., 2020; Sumaiya Thaseen and Aswani Kumar, 2017), among others]. This research conducted an unbiased comparison of feature selection methods to determine the most effective FS method and ML algorithm for building energy use prediction.

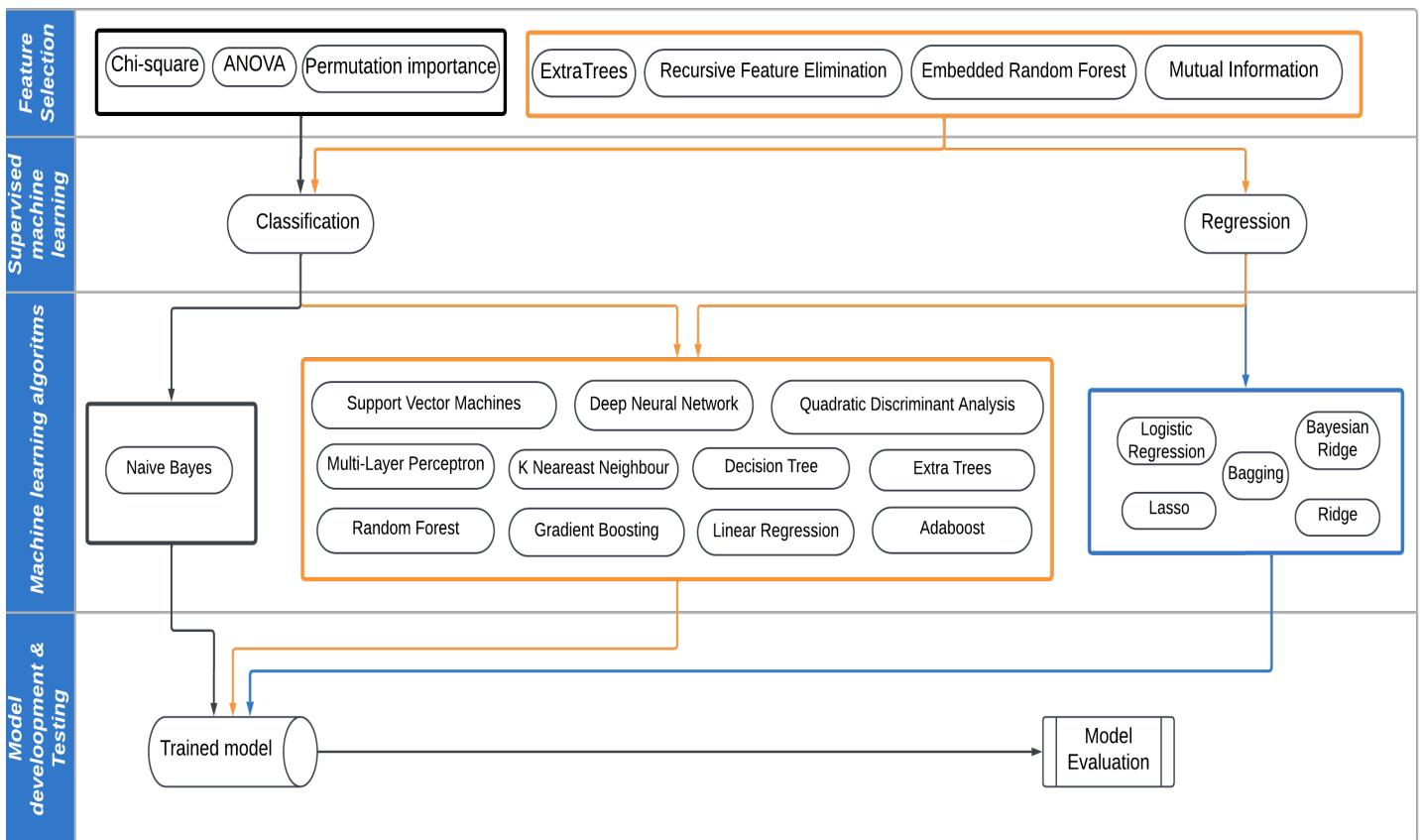


Figure 6.1: The framework of this analysis.

To address hypothesis H0, the research will apply different feature selection methods to analyse the identification of openings as the most relevant variables. Additionally, this research will perform a correlation analysis to determine the relationship between openings (windows and doors) and building energy consumption.

6.4 Feature Selection Methods

There are different of feature selection methods namely: Filter method, Chi-square method and Wrapper method.

6.4.1 Filters

Filter method is a type of feature selection method that provides more generality as they are independent of the chosen algorithm. They are fast in execution and computationally inexpensive in comparison to the wrapper and embedded techniques (Asir et al., 2016). Each of the filter methods employed in this research is explained below.

- **Chi-square:** This is a type of univariate filter FS test that calculates the deviation from the projected distribution considering the feature occurrence is independent of the label values (Sumaiya Thaseen and Aswani Kumar, 2017). Like any univariate method, chi-square is calculated between each feature and target or dependent feature, and then the presence of a correlation between them is detected. Subsequently, a low score is assigned if the target variable is independent of the feature while if the target variable is dependent on the feature, the feature is considered important (Effrosynidis and Arampatzis, 2021). Therefore, the higher the Chi-Square value, the more relevant the feature.
- **Mutual Information(MI):** This is a type of filter FS method proposed by (Battiti, 1994). It is also known as information gain. Mutual information aims to amplify the relevance between the input and output features and decrease the redundancy of the chosen features (Amiri et al., 2011). Subsequently, if the information gain of a feature is high, it is considered relevant. However, mutual information does not identify redundant features, because the features are chosen in a univariate way (Effrosynidis and Arampatzis, 2021).
- **ANOVA:** This is a type of univariate filter-based technique that utilizes variance to detect the separability of each feature between classes (Ding et al., 2014).

6.4.2 Wrappers

The wrapper method utilizes algorithms to assess the produced subsets by using the searching technique, making it more computationally complex (Asir et al., 2016). The Wrapper methods employed in this research are explained below.

- **Recursive Feature Elimination (RFE):** This is a type of multivariate wrapper technique that utilizes the decision tree classifier for training the model repeatedly using the existing features. The least important features are then eliminated using the weight of the algorithm as a ranking measure (Seijo-Pardo et al., 2019).

6.4.3 Embedded

These methods perform feature selection during the model training process itself, integrating it with the learning algorithm (Ang et al., 2016). Each of the Embedded methods employed in this research is explained below.

- **Embedded Random Forest (ERF):** This is an embedded method using the random forest algorithm. The significance of each feature is calculated by conducting random permutations of features in the out-of-bag set and measuring the increase in misclassification level in comparison to the out-of-bag set in the default state (Effrosynidis and Arampatzis, 2021).
- **ExtraTrees(ET):** ExtraTrees abbreviated as extremely randomized trees, is a form of random forest which performs randomization at every step for the selection of an optimal split. In contrast to random forests where the split features are centred on a grade, ExtraTrees implements a split measure of the random and considers the whole training set (Sharma et al., 2019).

6.5 Result and Discussion

In conducting a comprehensive data analysis, this section will perform Feature Selection Analysis for both classification and regression tasks to identify the most relevant predictors. For classification, techniques such as Chi-Square, ANOVA, Permutation Importance, Random Forest, Extra trees, Recursive Feature Elimination and mutual information will be employed to identify features that contribute significantly to the prediction of categorical outcomes. Similarly, for regression tasks, methods like Random Forest, Extra trees, Recursive Feature Elimination and mutual information will be utilized to identify key features influencing

continuous target variables. Complementing feature selection, Correlation Analysis will be conducted to understand the relationships between features and target variables. In both classification and regression, the Pearson correlation coefficient will be used to assess the linear relationships, ensuring a strong understanding of feature connections and dependencies in the dataset.

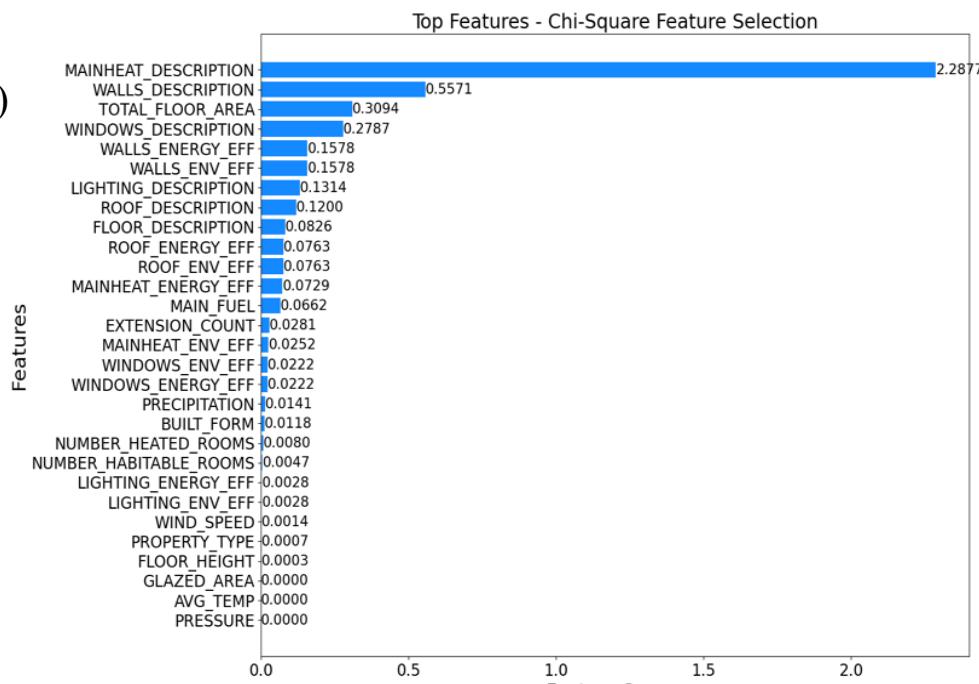
6.5.1 Classification

In the feature selection analysis for classification, the primary objective is to identify the most relevant features that contribute to the accurate classification of categorical target variables. The categorical variable utilized is the energy efficiency rating (see section 5.5.3 for details). This involves systematically evaluating and selecting a subset of features from the dataset to improve model performance and reduce overfitting. Various methods such as Chi-Square, ANOVA, Permutation Importance, Random Forest, Extra trees, Recursive Feature Elimination and mutual information are employed to assess feature importance.

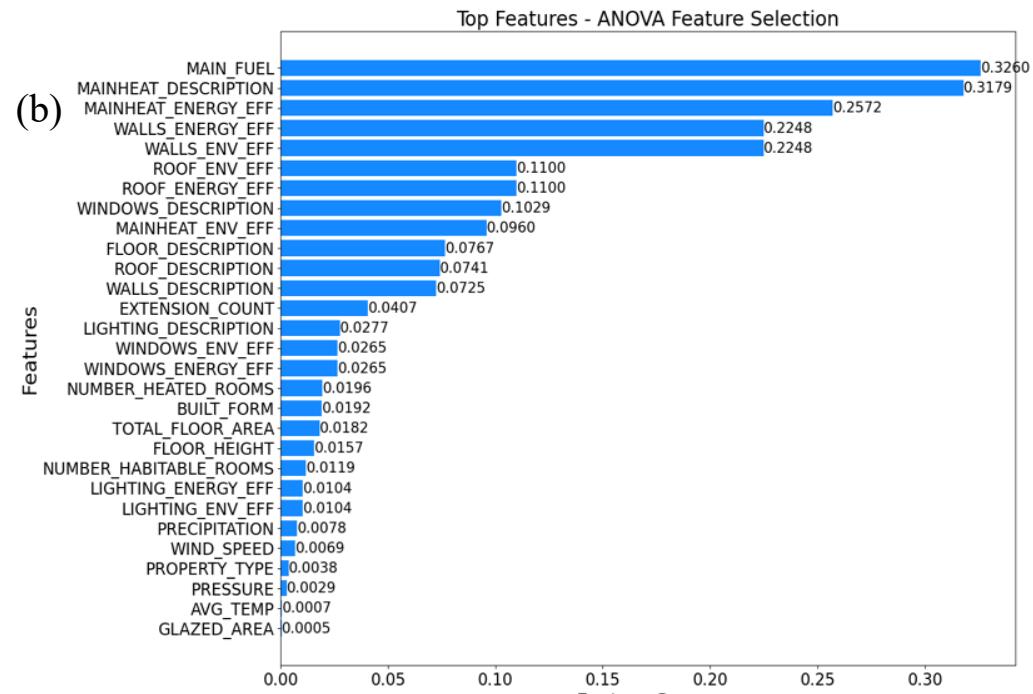
6.5.1.1 Feature Selection Analysis

Figure 6.2 below shows the charts for feature rank based on the importance scores. The seven feature selection methods utilized for classification are displayed below.

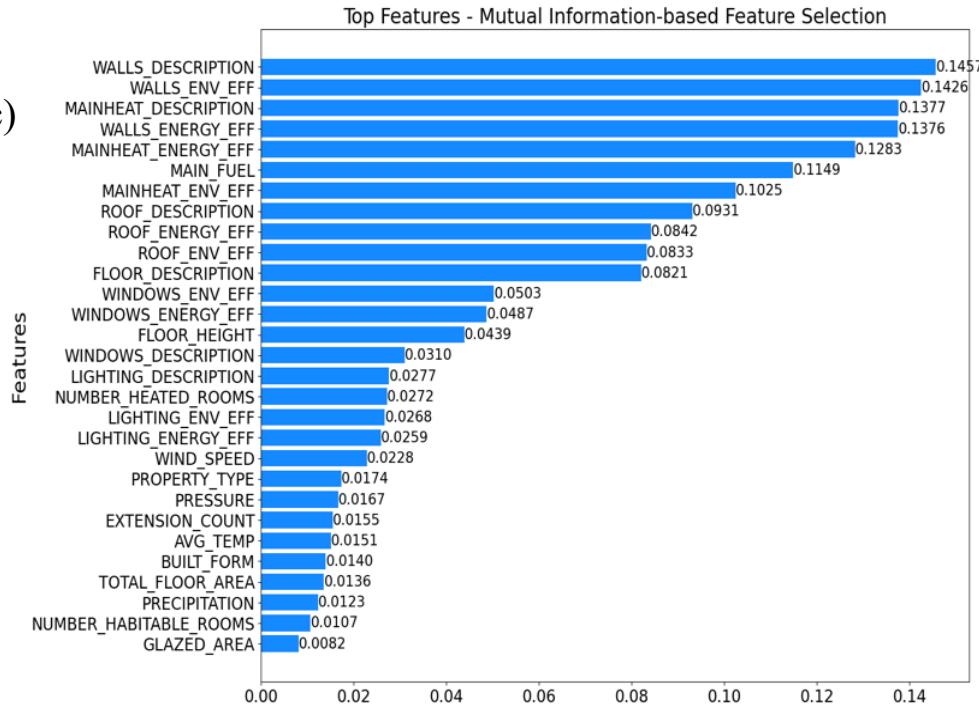
(a)



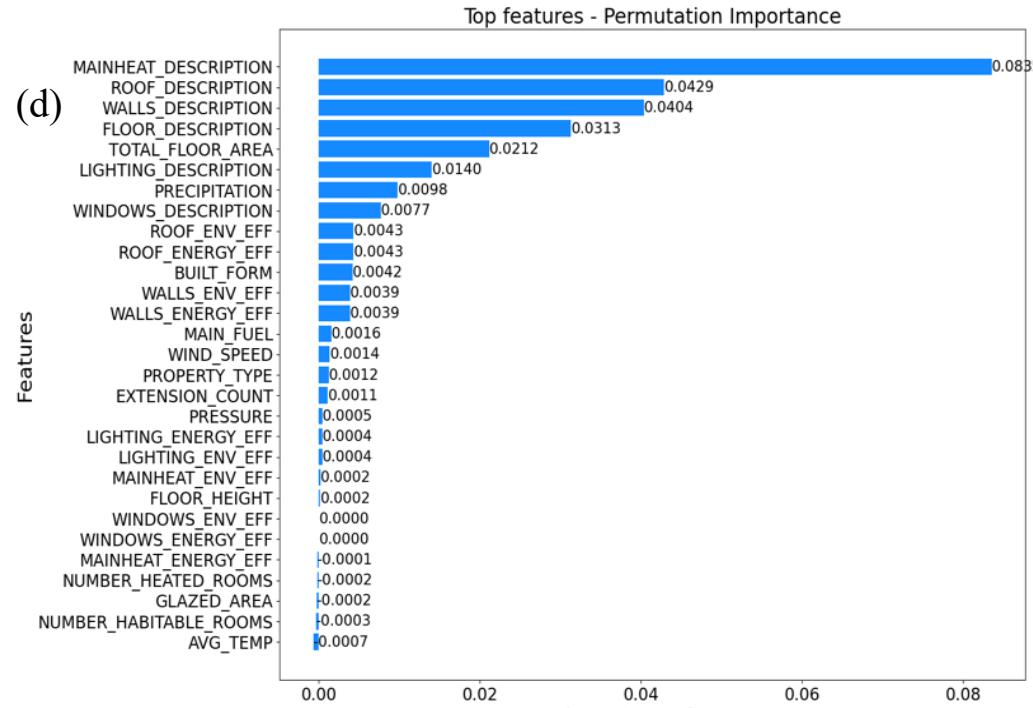
(b)



(c)



(d)



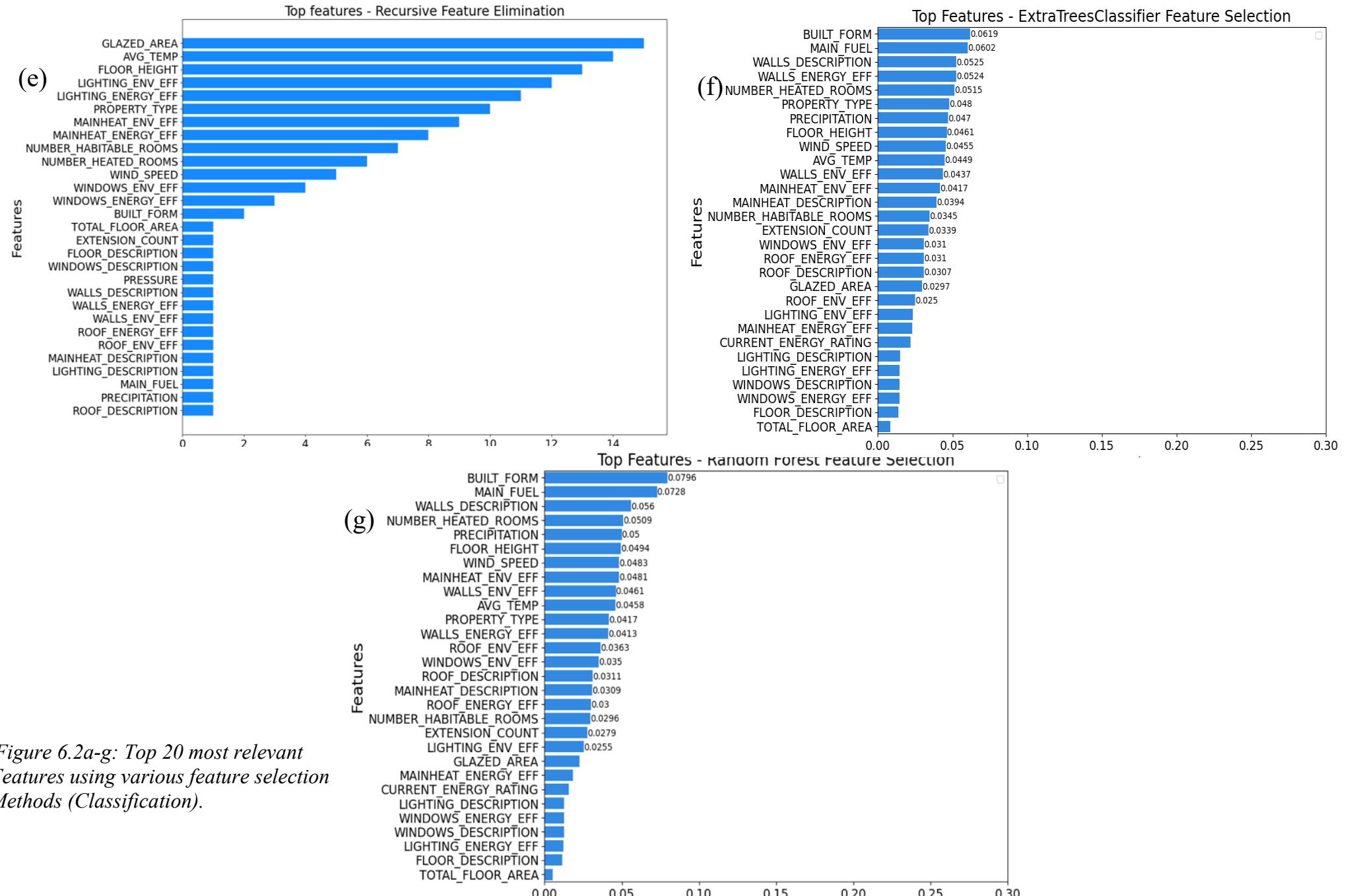


Figure 6.2a-g: Top 20 most relevant Features using various feature selection Methods (Classification).

Following the application of the feature selection methods, the highest-ranked features were selected as shown in Table 6.1 below. More specifically these are features that are considered important and hold values greater than 0.02(Pinheiro et al., 2020; Yin et al., 2023). Amongst these features, the most commonly ranked highly among the feature selection methods are wall description, roof description, and mainheat description. These are important features of a building and have an effect on the energy use of the building. For instance, walls are one of the most essential features of a building, as the selection of the most appropriate type of walls has an effect on the energy performance of a building (i.e., the thickness of the wall will reduce the use of heating in the winter season) (Marwan, 2020). Also, the roof has been acknowledged as the most essential structural component of the building with the potential to significantly reduce energy consumption of the building when optimized properly(Aboelata, 2021; Florides et al., 2002). However, pressure was not selected as a relevant feature across all feature selection methods. Table 6.1 shows the marked features which indicate features with importance values greater than 0.02 for each feature selection algorithm.

Table 6.1: Rank of each feature selected using various feature selection methods

	Features	Chi-square	Mutual Information	ANOVA	Permutation importance	RFE	RF	ExtraTrees
V1	PROPERTY_TYPE					X	X	X
V2	BUILT FORM					X	X	X
V3	TOTAL FLOOR AREA	X			X			
V4	GLAZED AREA					X		X
V5	EXTENSION COUNT	X		X			X	X
V6	NUMBER HABITABLE ROOMS					X	X	X
V7	NUMBER HEATED ROOMS		X			X	X	X
V8	FLOOR DESCRIPTION	X	X	X	X			
V9	WINDOWS DESCRIPTION	X	X	X				
V10	WINDOWS ENERGY EFF	X	X	X		X		
V11	WINDOWS ENV EFF	X	X	X		X	X	X
V12	WALLS DESCRIPTION	X	X	X	X		X	X
V13	WALLS ENERGY EFF	X	X	X			X	X
V14	WALLS ENV EFF	X	X	X			X	X
V15	ROOF DESCRIPTION	X	X	X	X		X	X
V16	ROOF ENERGY EFF	X	X	X			X	X
V17	ROOF ENV EFF	X	X	X			X	X
V18	MAINHEAT DESCRIPTION	X	X	X	X		X	X
V19	MAINHEAT ENERGY EFF	X	X	X		X		
V20	MAINHEAT ENV EFF	X	X	X		X	X	X
V21	LIGHTING DESCRIPTION	X	X	X				
V22	LIGHTING ENERGY EFF		X			X		
V23	LIGHTING ENV EFF		X			X	X	
V24	MAIN FUEL	X	X	X			X	X
V25	FLOOR HEIGHT		X			X	X	X
V26	AVG TEMP					X	X	X
V27	PRECIPITATION						X	X
V28	WIND SPEED		X			X	X	✓
V29	PRESSURE							

6.5.1.2 Correlation Analysis (Classification)

Result from the application of feature selection show that some weather-related features (i.e., wind speed, outdoor temperature among others) and window related feature (i.e., window description, window environmental efficiency among other) were identified to have comparatively lower ranking of importance as shown in Table 6.1 and Figure 6.2a-g. Nonetheless, low importance features do not directly mean the features are irrelevant to the target variable. Thus, some features significantly correlated with each other and the result for the evaluation of this correlation is provided in Figure 6.3 below. The correlation matrix plot shows the correlation between features (building and meteorological features) and target (energy efficiency rating).

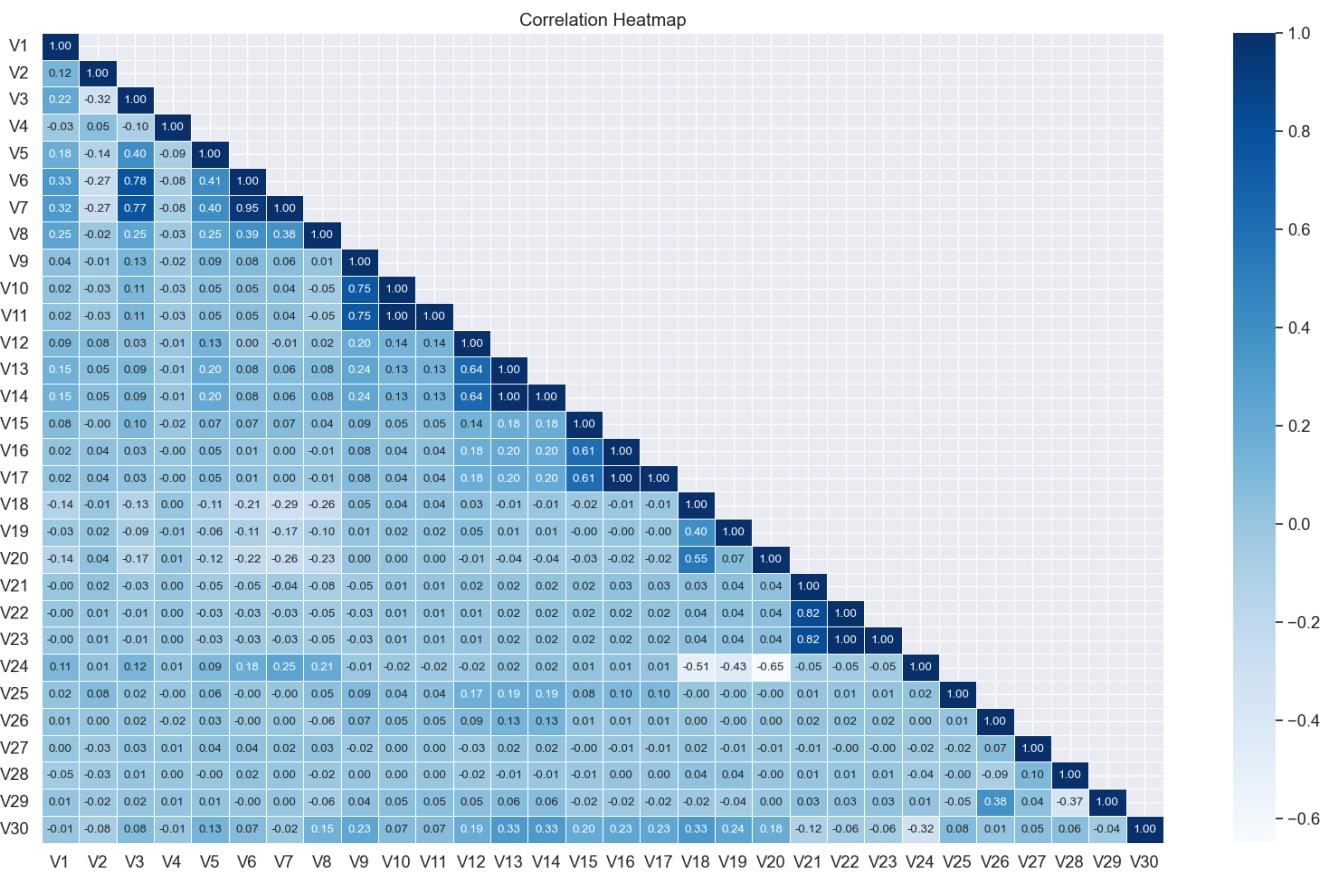


Figure 6.3.: Correlation between variables (classification)

As shown in figure 6.3, the diagonal line showing one represents the correlation of the features with themselves. The correlation values close to one represents a strong positive correlation between two features. The result shows that weather and window features had no significant correlation with other features except related features such as window environmental

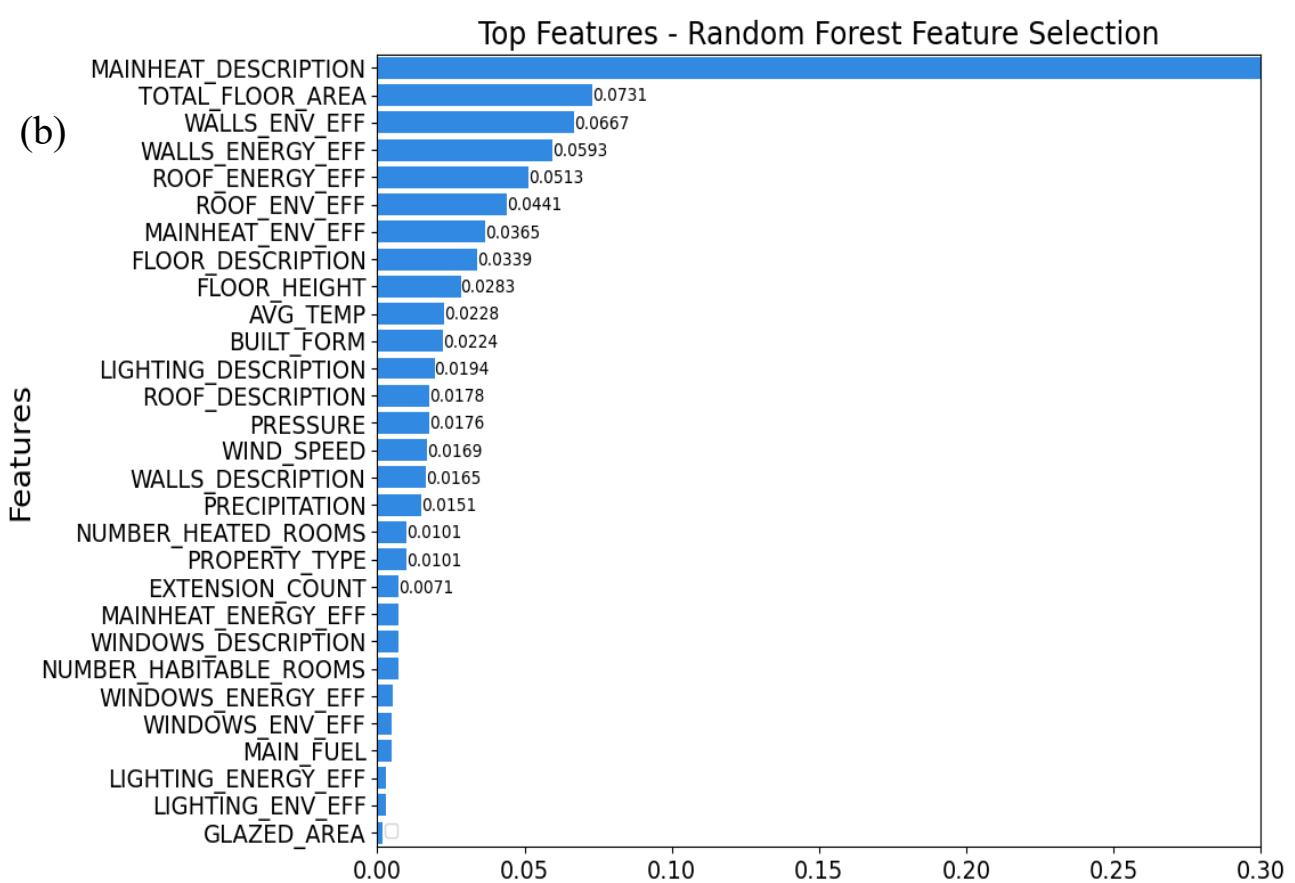
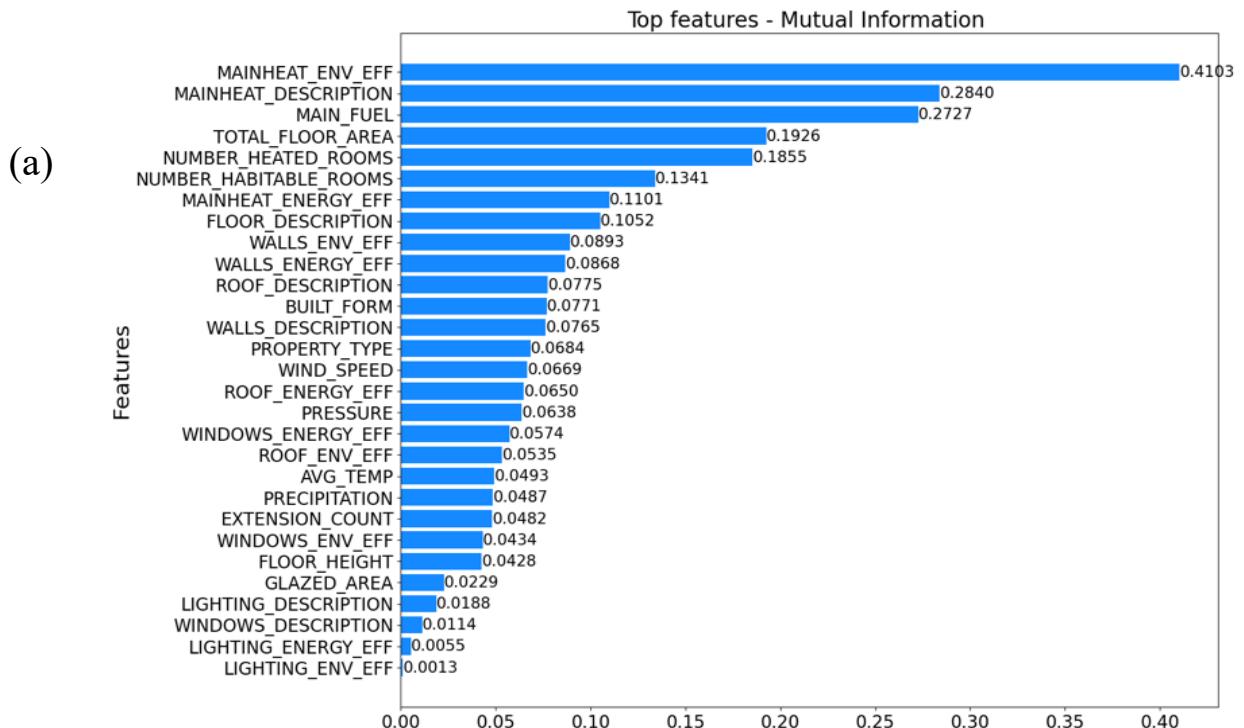
efficiency, window energy efficiency. However, number of heated rooms has a very strong correlation to a number of habitable rooms with a correlation value of 0.95 and they both have a strong positive correlation of 0.75 with the total floor area. Thus, this analysis supports the hypothesis that openings (such as windows and doors) do not have the most significant effect on building energy consumption (See section 6.5.5). For a clear and comprehensive confirmation, similar analysis was conducted for regression task according to the framework in figure 6.1.

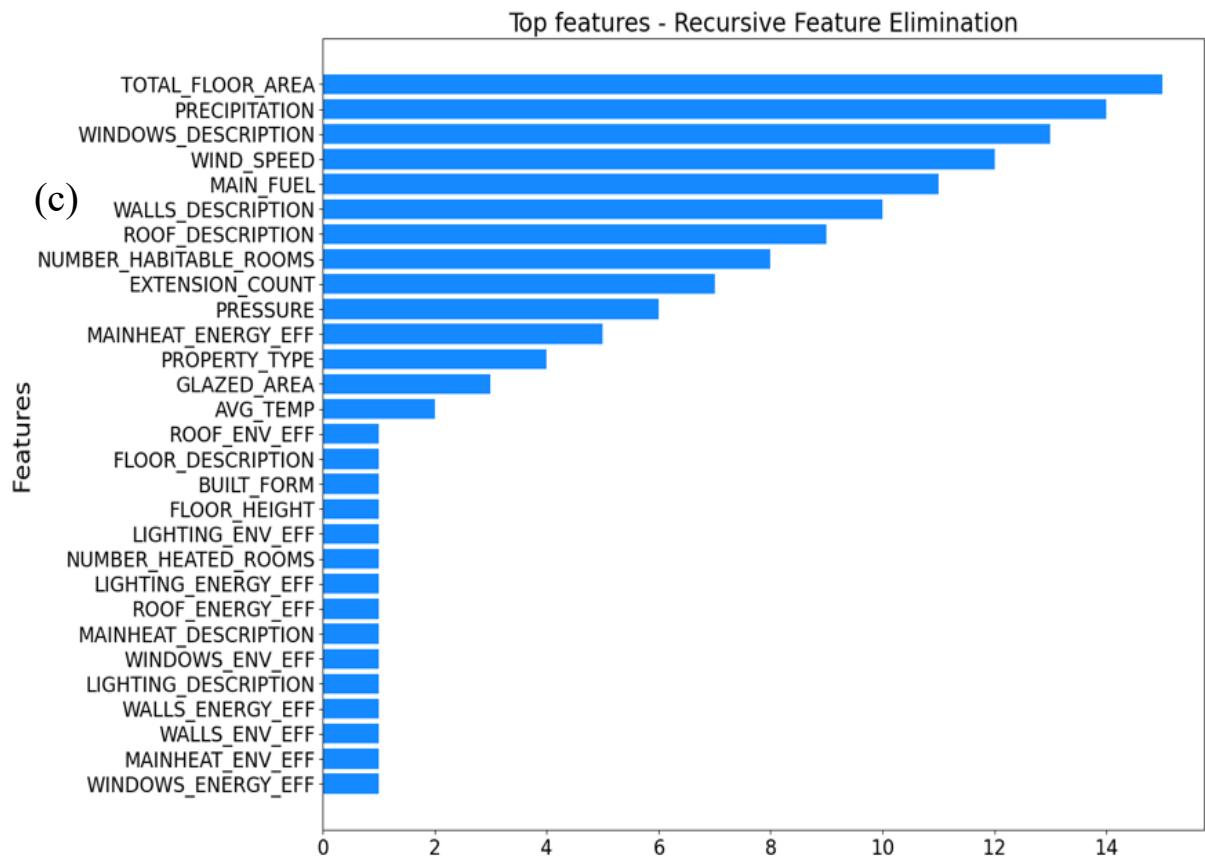
6.5.2 Regression

Furthermore, to avoid bias and to make a clear and unambiguous conclusion, this research also applies feature selection methods for regression where energy consumption values are employed as the target feature. This will corroborate or invalidate the result of feature selection methods. Considering it remains a prevailing argument in research that feature selection is more effective in classification than regression and vice versa. While some studies have argued this is true (Jović et al., 2015; Kumar, 2014), several studies have employed feature selection for regression prediction and achieved outstanding performance(Ahmad et al., 2017b; Fan et al., 2014a; Kolter and Ferreira, 2011; Paudel et al., 2017). In the feature selection analysis for regression, the goal is to identify the most relevant features that contribute to the accurate prediction of target continuous variable. The continuous variable utilized is the energy consumption value (see section 5.5.3 for details). Various methods such as Recursive Feature Elimination, Random Forest, Extra trees, and mutual information are employed to assess feature importance.

6.5.2.1 Feature Selection Analysis

Subsequently, feature selection was applied for the regression task. Only MI, RFE, RF, and ET feature selection methods were used for regression because chi-square, ANOVA among others are considered unsuitable for regression prediction tasks. This is because they are statistical tests used to determine significant associations between categorical variables. The result of the feature selection analysis is displayed in Figure 6.4a-d.





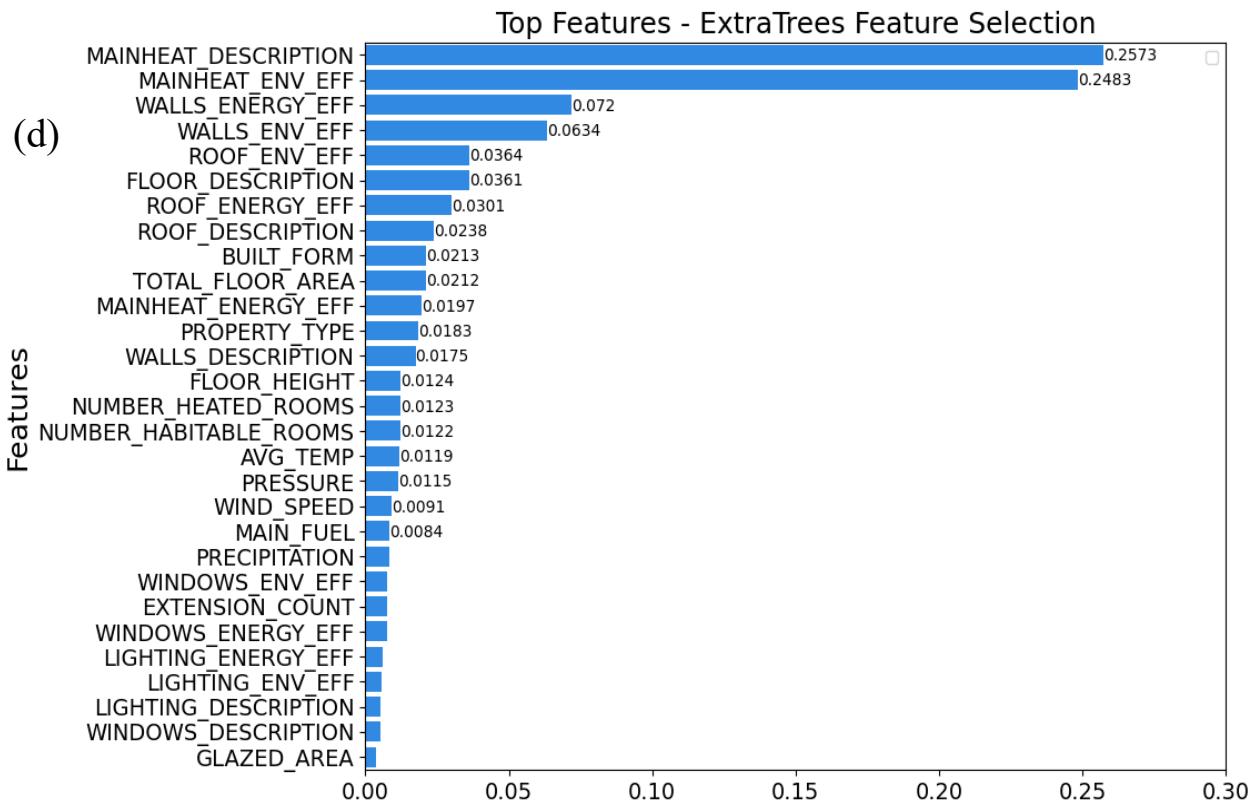


Figure 6.4a-d: Top 20 most relevant features using various feature selection methods(regression).

Figure 6.4a-6.4d shows the ranking most relevant features using five feature selection method. Following the application of the feature selection methods, the highest ranked features that were above 0.2 were selected as shown in Table 6.2 below. Amongst these, the most common highest ranked features across all feature selection methods are Total floor area and floor description. PCA selected floor, roof and weather-related feature (precipitation) as the most relevant features, however RF and MI selected main heat description as the most relevant.

Additionally, window description was not selected as relevant feature across all feature selection methods

Table 6.2: Rank of each feature selected using various feature selection methods

	Features	Mutual Information	RFE	Random Forest	ExtraTrees
V1	PROPERTY_TYPE	X	X		
V2	BUILT_FORM	X		X	X
V3	TOTAL_FLOOR_AREA	X	X	X	X
V4	GLAZED_AREA	X	X		
V5	EXTENSION_COUNT	X	X		
V6	NUMBER_HABITABLE_ROOMS	X	X		
V7	NUMBER_HEATED_ROOMS	X			

V8	FLOOR DESCRIPTION	X		X	X
V9	WINDOWS DESCRIPTION	X	X		
V10	WINDOWS ENERGY EFF	X			
V11	WINDOWS ENV EFF	X			
V12	WALLS DESCRIPTION	.	X		
V13	WALLS ENERGY EFF	X		X	X
V14	WALLS ENV EFF	X		X	X
V15	ROOF DESCRIPTION	X	X		X
V16	ROOF ENERGY EFF	X		X	X
V17	ROOF ENV EFF	X		X	X
V18	MAINHEAT DESCRIPTION	X		X	X
V19	MAINHEAT ENERGY EFF	X	X		
V20	MAINHEAT ENV EFF	X		X	X
V21	LIGHTING DESCRIPTION				
V22	LIGHTING ENERGY EFF				
V23	LIGHTING ENV EFF				
V24	MAIN FUEL	X	X		
V25	FLOOR HEIGHT	X		X	
V26	AVG TEMP	X	X	X	
V27	PRECIPITATION	X	X		
V28	WIND SPEED	X	X		
V29	PRESSURE	X	X		

6.5.2.2 Correlation Analysis

Subsequently, correlation analysis was conducted to identify features that are significantly correlated with the target feature. As displayed in Figure 6.3 above, the result from the correlation analysis of the classification task shows that none of the features have a strong correlation with the target variables (energy efficiency rating). However, Figure 6.5. shows that two features namely main heat description and main heat environmental efficiency have a strong correlation with the target variables (energy consumption values). The result shows that weather and window features had no significant correlation with other independent features and the target feature.

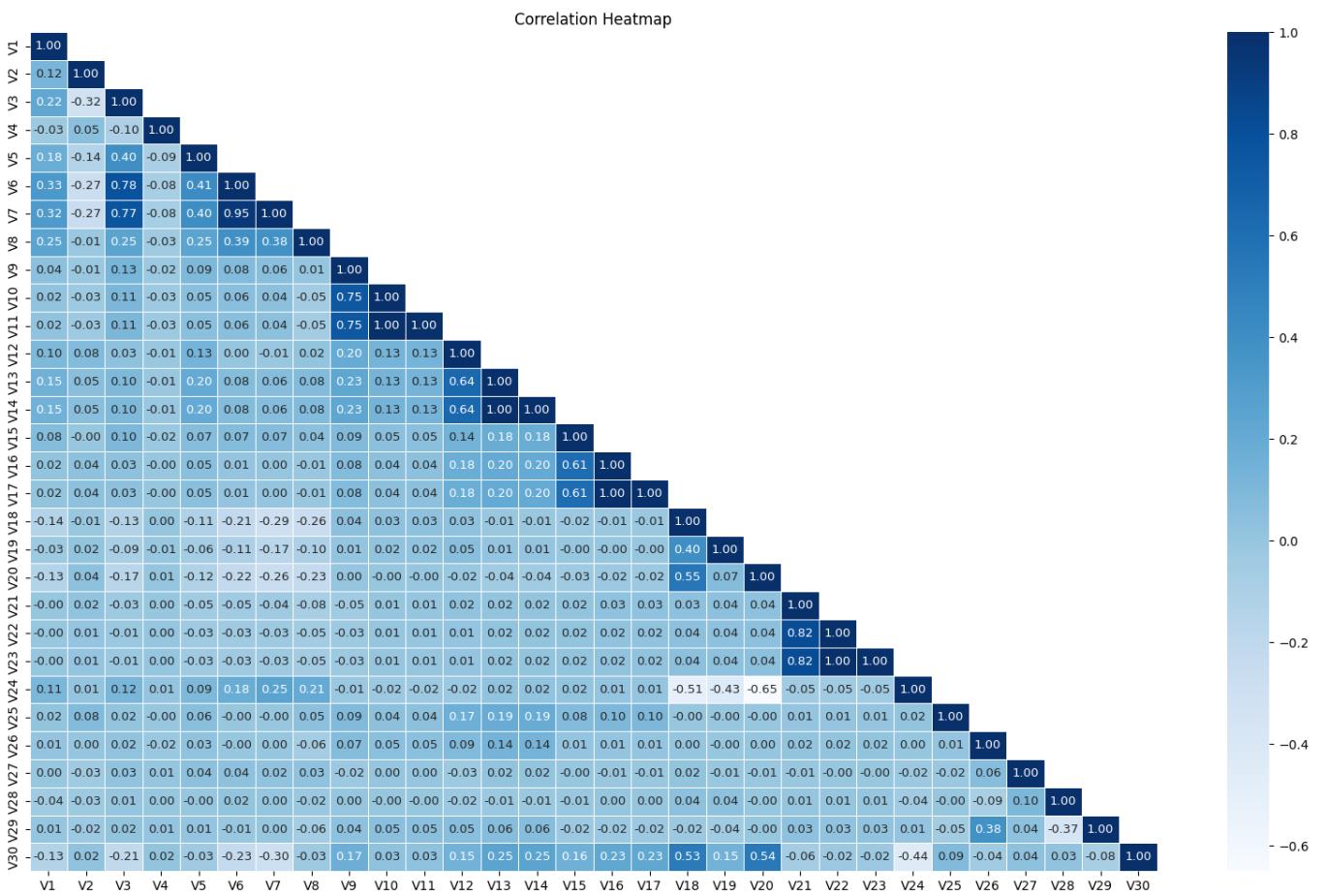


Figure 6.5: Correlation between variables (regression)

6.5.5 H₀: Openings (such as windows and doors) have the most Significant Effect on Building Energy Consumption.

Previous research finds that openings (such as windows) are the weakest component of the building envelope, accountable for a significant amount of heating and cooling energy consumption(M. Alwetaishi and Benjeddou, 2021; Chiesa et al., 2019; Yoshino et al., 2017). However, some other research proclaims that outdoor temperature has the greatest effect on energy consumption in buildings (Ihara et al., 2015; Park et al., 2020; Rouleau et al., 2018). Hence, it is proffered that proper insulation of walls, and roof, among others have the most significant effect on the heating or cooling energy consumption of the building depending on the weather condition (L. Y. Zhang et al., 2017). The underlying null and alternative hypotheses are stated as follows:

- **Null hypothesis H₀:** Openings (such as windows and doors) have the most significant effect on building energy consumption.
- **Alternative hypothesis H_A:** Openings (such as windows and doors) do not have the most significant effect on building energy consumption.

The result of the feature selection analysis (see section 6.5.1.1 and 6.5.2.1) shows that building components such as walls, and roofs were rated some of the most relevant features among all the feature selection methods employed. Therefore, the null hypothesis(H_0) is rejected. The outcome of this investigation supports the alternative hypothesis, as explicitly stated by Zhang et al., (2017). Openings such as Windows emanated as one of the top relevant features in 3 out of the 7 feature selection methods utilized. Although this indicates that windows do have an effect on the energy consumption of a building, it does not have the most significant effect on the energy consumption of a building. Using the mutual information-based feature selection method, wall description was ranked the top(1st) most relevant feature while window description was ranked the 15th most relevant feature as shown in Figure 6.2c. Likewise, wall description was ranked the 3rd most relevant feature using random forest, extra trees and permutation importance while window description was ranked 8th, 26th and 26th most relevant feature using permutation importance, random forest and extra trees respectively. Evidently, weather data has effect on energy consumption, however, weather features such as wind speed, pressure, and temperature among others were selected among the top most relevant features in only a few feature selection methods namely Recursive Feature Elimination (RFE), mutual information, Embedded Random Forest and ExtraTrees. Notwithstanding, RFE selected average outdoor temperature as the 2nd most relevant feature that has an effect on the target variable while wind speed was selected as the 5th most relevant using the random forest feature selection method as shown in figure 6.2e&6.2f. Thus, it is noted that this is not enough to make a clear and unbiased conclusion, hence this research further tests this hypothesis by conducting a correlation analysis (see section 6.5.2.1).

To further exemplify the impact of the features on energy consumption, the Ordinary Least Squares (OLS) regression analysis was implemented. OLS helps to understand and quantify the relationships between the different independent variables and the dependent variables, ('ENERGY_CONSUMPTION_CURRENT[V30]'). Thus, this analysis offers a broad understanding of the influential factors on energy consumption. OLS is a statistical method employed for modelling the linear connection between a dependent variable and one or more independent variables. The model result provides insights into the overall fit and significance of the model, with reported R-squared, and F-statistic. The results also present the coefficients, standard errors, t-values, and p-values for all independent variables. These values help evaluate the strength and significance of the relationships between the variables. Additionally, statistics such as Omnibus, Durbin-Watson, and Jarque-Bera provide insights into model assumptions

and goodness of fit. Also, a condition number is utilized to detect multicollinearity. Condition number with high values suggests potential issues.

Figure 6.6 includes information about the significance level of each variable. The significance level is typically measured using the p-value associated with each coefficient in the regression model. A low p-value (typically below a significance threshold of 0.05) denotes that the variable is statistically significant.

OLS Regression Results						
Dep. Variable:	ENERGY_CONSUMPTION_CURRENT	R-squared:	0.581			
Model:	OLS	Adj. R-squared:	0.581			
Method:	Least Squares	F-statistic:	5675.			
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	0.00			
Time:	14:28:28	Log-Likelihood:	-5.8527e+05			
No. Observations:	102148	AIC:	1.171e+06			
Df Residuals:	102122	BIC:	1.171e+06			
Df Model:	25					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	6311.1493	267.601	23.584	0.000	5786.655	6835.644
PROPERTY_TYPE	-9.9491	0.370	-26.865	0.000	-10.675	-9.223
BUILT_FORM	-3.5331	0.138	-25.546	0.000	-3.804	-3.262
TOTAL_FLOOR_AREA	-0.3405	0.008	-44.145	0.000	-0.356	-0.325
GLAZED_AREA	1.2736	0.417	3.052	0.002	0.456	2.092
EXTENSION_COUNT	4.2968	0.324	13.265	0.000	3.662	4.932
NUMBER_HABITABLE_ROOMS	10.0501	0.460	21.827	0.000	9.148	10.953
NUMBER_HEATED_ROOMS	-15.6521	0.449	-34.889	0.000	-16.531	-14.773
FLOOR_DESCRIPTION	4.7737	0.065	73.040	0.000	4.646	4.902
WINDOWS_DESCRIPTION	8.9040	0.120	74.139	0.000	8.669	9.139
WINDOWS_ENERGY_EFF	-10.9143	0.210	-52.030	0.000	-11.325	-10.503
WINDOWS_ENV_EFF	-10.9143	0.210	-52.030	0.000	-11.325	-10.503
WALLS_DESCRIPTION	-0.5766	0.026	-22.141	0.000	-0.628	-0.526
WALLS_ENERGY_EFF	9.5055	0.106	89.830	0.000	9.298	9.713
WALLS_ENV_EFF	9.5055	0.106	89.830	0.000	9.298	9.713
ROOF_DESCRIPTION	0.5591	0.039	14.313	0.000	0.483	0.636
ROOF_ENERGY_EFF	7.1480	0.107	66.637	0.000	6.938	7.358
ROOF_ENV_EFF	7.1480	0.107	66.637	0.000	6.938	7.358
MAINHEAT_DESCRIPTION	1.3167	0.011	115.314	0.000	1.294	1.339
MAINHEAT_ENERGY_EFF	-8.9331	0.383	-23.307	0.000	-9.684	-8.182
MAINHEAT_ENV_EFF	46.7804	0.487	96.131	0.000	45.827	47.734
LIGHTING_DESCRIPTION	-0.4994	0.016	-31.489	0.000	-0.530	-0.468
LIGHTING_ENERGY_EFF	2.9078	0.194	14.991	0.000	2.528	3.288
LIGHTING_ENV_EFF	2.9078	0.194	14.991	0.000	2.528	3.288
MAIN_FUEL	-4.4770	0.141	-31.666	0.000	-4.754	-4.200
FLOOR_HEIGHT	6.5623	0.924	7.100	0.000	4.751	8.374
AVG_TEMP	-7.4453	0.314	-23.685	0.000	-8.061	-6.829
PRECIPITATION	0.2892	0.017	17.020	0.000	0.256	0.322
WIND_SPEED	-0.7753	0.113	-6.874	0.000	-0.996	-0.554
PRESSURE	-5.9485	0.264	-22.535	0.000	-6.466	-5.431
Omnibus:	25046.043	Durbin-Watson:	1.992			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	155865.314			
Skew:	1.035	Prob(JB):	0.00			
Kurtosis:	8.686	Cond. No.	1.00e+16			

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.08e-21. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Figure 6.6: OLS Regression analysis

V18 has the highest t-value of 115.3. In this context, a higher absolute t-value signifies an extremely strong and statistically significant relationship between V18 and V30 (dependent variable). The strong positive effect of V18 implies that the variation of the main heating system and its relative source (i.e., 'Boiler and radiators [mains gas]', 'Electric storage heaters', 'Boiler and radiators, LPG' among others) significantly contribute to higher energy consumption in buildings. It is proffered that developers should be scrupulous in the selection

of the specific characteristics or type of the main heating system for the building when assessing and managing energy efficiency. The result of this analysis conveys that changes or improvements in the main heating system could have a considerable impact on the overall building energy consumption patterns.

Furthermore, coefficients represent the projected change of V30 for a one-unit change in the corresponding variable, all things being equal. For instance, V3 has a positive relationship and high statistical significance ($\text{coef} = -0.3405$, $t = -44.14$, $p < .0001$). This indicates that a one-unit or square meter increase in total floor area is related to a decrease of 0.3405 units in energy consumption. Likewise, V1 shows a significant positive effect ($\text{coef} = -9.9491$, $t = -26.87$, $p < .0001$). This suggests that the type of property significantly influences the increased energy consumption of a building. Conversely, V6 elicited a coefficient of 10.0501 which conveys that an increase in the number of habitable rooms corresponds to an increase of over 10.05 units in energy consumption. This suggests that buildings with a larger number of habitable rooms are likely to have higher energy consumption levels. In building energy consumption analysis, a broad understanding of both positive and negative relationships is imperative for identifying features contributing to efficiency or increased energy consumption. While the analysis does reveal influencing factors of energy consumption, it also unveils areas of concern that could potentially lead to issues in model development, necessitating further investigation. For instance, the diagnostic information in the result raises concerns about potential multicollinearity problems, given the small eigenvalue.

There are different strategies which can be used to address multicollinearity in regression tasks such as correlation analysis and feature selection, which are both implemented in this research. Multicollinearity is a condition that occurs when two or more independent variables are highly correlated with each other (Chan et al., 2022). The examination of the correlation matrix of the independent variables to identify two or more features with high correlations to each other and one of them can be removed. Contrarily, feature selection techniques can be employed to select a subset of relevant independent variables, thereby minimizing multicollinearity.

6.8 Chapter Summary

This chapter presents a thorough examination of feature selection methods and their impact on machine learning models for building energy consumption prediction. It delivers a comprehensive methodology section outlining the research approach. Various feature selection methods including Filters, Wrappers, Embedded, and Correlation Analysis are discussed. This chapter addressed one hypothesis of this research(H0).

CHAPTER SEVEN

7.0 DEVELOPMENT OF AI/ML BUILDING ENERGY CONSUMPTION PREDICTION MODEL

7.1 Chapter Overview

This chapter investigates the impact of data size on the performance of building energy consumption prediction models. The identification of the optimum data size is required to achieve satisfactory model performance. This chapter examines the effect of data size in both classification and regression prediction model performance and delivers the theoretical and practical implications of such analysis. Additionally, this chapter will employ big data analytics to analyse the various model performance based on accuracy and computational efficiency. These measures include Root Mean Squared Error (RMSE) Mean Absolute Error (MAE) and Accuracy among others. Additionally, this chapter will comparatively analyse the impact of model performance using 20% - 100% data availability. This chapter will address nine key hypotheses (**H1 to H9**).

7.2 Background on Energy Prediction Model

In the quest to identify the optimal algorithm for annual energy consumption, it is important to acknowledge that the accuracy result of the model applied on different datasets in different studies in literature is not directly comparable because different data and different situations will produce different results (Demsar, 2006). Hence, it cannot be concluded that an algorithm is better than the others unless it has been analysed and compared on the same dataset. It is well noted that the size of the data is important in developing BEP models (Liang Zhang et al., 2021). AI-based tools rely on historical data to learn energy consumption patterns (Somu et al., 2021). However, some tools tend to thrive better in small datasets such as SVM (Mat Daut et al., 2017) while some require large datasets to produce good results such as DNN (Ngarambe et al., 2020). Feng and Zhang, (2020) emphasized the consensus that the simple application of a single model for prediction is not suitable as the performance of a prediction model is considerably influenced by the data and no single model in its stand-alone form can be noted the best amongst all models. DNN is recognized for producing good performance in a large dataset and this has been analysed and performance was emphasized in the study by Sadeghi et al., 2020. Amber et al., (2018) stipulated that DNN models are not as favourable in studies with a limited amount of data, thus its performance relies heavily on a large amount of data.

Due to the small amount of data with high dimensions used in the study by Guo et al., (2020b), RF produced the worst tools while a hybrid model, boruta feature selection applied with SVM produced the best prediction performance among others.

Owing to prominence, ANN is the most predominant choice for energy use prediction in buildings (Ahmad et al., 2017b; J.-H. Kim et al., 2020). Notwithstanding its drawbacks such as high computational cost and lack of transparency (Alaka et al., 2018; Mat Daut et al., 2017; Sadeghi et al., 2020), numerous studies have randomly employed ANN tool due to its acceptance in the field of energy prediction (Almalaq and Zhang, 2019; Hwang et al., 2020; Izidio et al., 2021; Koukaras et al., 2021). ANN is fairly recognised for its production of good outcomes in the availability of large data size to train the model (Alaka et al., 2018; Bourhnane et al., 2020; Olu-Ajayi et al., 2022b) However, ANN and other fairly common data-driven tools (i.e., RF, LR) have been utilized and compared in various studies using small data size (Bouktif et al., 2020; Groß et al., 2021). More recently, SVM has emerged as one of the most utilised data-driven tools based on its capacity to produce good outcomes regardless of the data size (Mat Daut et al., 2017; Olu-Ajayi et al., 2021). However, a drawback of SVM is its large requirements and low computational efficiency (Zhang et al., 2005). Various comparative analysis of ANN and SVM have been conducted and some studies concluded that SVM performs better than ANN while some conclude otherwise(Chammas et al., 2019; Feng and Zhang, 2020; Li and Yao, 2020), which may be subject to the size of data employed.

Moreover, within the field of machine learning, it is a widely accepted hypothesis that the predictive model performance is positively impacted by the size of the dataset employed [i.e. the larger the data, the more accurate the result (Goyal et al., 2020; Kabir, 2020; Kaur and Gupta, 2017; Lee et al., 2011; Olu-Ajayi et al., 2022a)]. Therefore, this chapter investigated the effect of sample size on the performance of ML models. There are very few studies that have conducted a comparative analysis of the major algorithms on the same data to identify best-performing model in building energy prediction. Notably, there is a shortage of studies conducting a comparative analysis of major algorithms on datasets of various data sizes for building energy prediction. This investigation involves fixing the data employed for training of each model at different percentages — 20%, 40%, 60%, 80%, and 100%—to enable a clear and comprehensive comparison as shown in Figure 7.1 below. Therefore, the dissimilar quantities of training sets signify different types of data, for example, 20% and 40% represent the inadequate amount of data, and 60% represent good while 80% and 100% represent an adequate amount of data.

7.3 Statistical and Machine Learning Tools

Due to their prominence and systematic review, 17 statistical and machine learning tools were used to develop classification and regression models. For the classification model, 12 statistical and machine learning tools were employed, including 2 statistical (ST) methods and 10 machine learning(ML) techniques: Support Vector Machine (SVM), Gradient Boosting (GB), Random Forest (RF), Linear Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Naive Bayes (NB), AdaBoost, Extra Trees (ET), Multilayer Perceptron (MLP), Quadratic Discriminant Analysis (QDA), and Bagging. For regression tasks, 15 statistical and machine learning tools were utilized, comprising 5 statistical methods and 10 machine learning techniques: Support Vector Regression (SVR), Artificial Neural Networks (ANN), Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), AdaBoost, Bagging, Extra Trees (ET), Ridge Regression, Lasso Regression, Bayesian Ridge (BR), Least Squares Regression (LSR), and Deep Neural Networks (DNN). These methods are briefly explained below (see section 2.4 for more details).

- **(ST) Linear Regression(LR):** presents the association among variables by fitting a linear equation to the data (Fumo and Rafe Biswas, 2015). Linear regression proffers some advantages such as ease of use, interpretability and so on (Pino-Mejías et al., 2017). The linear fitting is conducted by making all but one predictor variable constant. The relationship between a predictor variable and a response variable does not suggest that the predictor variable causes the response variable, but that there is a significant relationship between the two variables.
- **(ST) Quadratic Discriminant Analysis:** This is a common method for supervised classification, which gaussian distribution in modelling the likelihood of each class, then utilizes posterior distribution to predict the class (Srivastava et al., 2007). The default parameters were utilized in the development of the Quadratic Discriminant Analysis.
- **(ST) Ridge Regression:** is a linear regression method utilized to mitigate multicollinearity and overfitting by adding an L2 penalty term to the loss function, which removes large coefficient values(Mohanty and Palai, 2023). This penalty term is proportional to the square of the magnitude of coefficients, promising smaller coefficients and reducing model complexity(Liqiang Zhang et al., 2021).

- **(ST) Lasso Regression:** This is a linear regression method that adds a penalty term to the loss function to prevent overfitting. It uses reduction to improve the predictive accuracy and interpretability of regression models(Q. Zhang et al., 2020).
- **(ST) Bayesian Ridge (BR):** This is a type of linear regression algorithm that uses probability distributions rather than point estimates in order to solve a regression problem (Demertzis et al., 2022).
- **(ST) Logistic Regression:** This is an algorithm that estimates the likelihood that a specified input belongs to a specific category based on its features, making it particularly more suitable for binary classification tasks like spam detection, and fraud detection among others(Konasani and Kadre, 2015).
- **(ML) Support Vector Machines:** This is a machine learning method also identified as Support Vector Classifier (SVC), it is recognised as one of the most accurate techniques amongst data mining algorithms (Wu et al., 2008). SVM has gained more attention, owing to its capability of effectively generating good solutions to non-linear problems in diverse sample sizes (Chen et al., 2022; Olu-Ajayi et al., 2023a; Hai-xiang Zhao and Magoulès, 2012b). issues and it can achieve optimum solutions worldwide (Wei et al., 2018).
- **(ML) Random Forest:** This is an ensemble technique developed based on the ensemble learning theory, which makes the learning of both simple and complicated problems achievable (Ahmad et al., 2017a); the RF algorithm often generates good performance using default parameters. Hence, the RF algorithm is obtaining increased recognition in the field of energy use prediction (Ahmad et al., 2017a; Y.-T. Chen et al., 2019; Cheng Fan et al., 2017; Z. Wang et al., 2018b).
- **(ML) Gradient Boosting:** This is a type of boosting method that develops models in phases, but it generalizes these by applying a random differentiable loss function (Flores and Keith, 2019). GB is one of the ML techniques that can be used for both classification and regression problems.
- **(ML) Decision Tree (DT):** is a method of utilizing a tree-like flowchart to partition data into groups. Decision Tree is an adaptable process that could advance with an

enlarged amount of training data (Domingos, 2012). In contrast to other data-driven methods, DT is easier to comprehend, and its application does not require complex computation knowledge. However, it often produces a major deviation of its predictions from actual results. DT is more suitable for forecasting categorical features than for estimating numerical variables (Yu et al., 2010).

- **(ML) K-Nearest Neighbour (kNN):** K-Nearest Neighbour (k-NN) algorithm is a non-parametric machine learning method that utilizes similarity or distance function d to predict outcomes based on the k nearest training examples in the feature space (Ortiz-Bejar et al., 2018). kNN algorithm is one of the common distance functions that works effectively on numerical data (Ali et al., 2019). However, KNN has yet to receive much attention in the field of building energy prediction
- **(ML) Adaboost:** This is the simplest of boosting methods, which can be used for solving classification problem (Rahul et al., 2018). Ada boost is known for its low computational time and its high detection speed (Aadithyan et al., 2020).
- **(ML) Bagging:** Bagging is an abbreviation of Bootstrap Aggregating, which is one of the most utilized and famous ensemble learning methods (Zeng et al., 2010). Bagging generates various classifiers in parallel and then ensembles them, so it chooses specific base classifier algorithms to train base classifiers on a random redistribution training dataset.
- **(ML) Naïve Bayes:** In the data mining and machine learning field, this ML method is considered one of the valuable classification methods due to its effectiveness. However, they fail in the assumption of conditional independence among the features (Jahromi and Taheri, 2017). The default parameters were utilized in the development of the NB.
- **(ML) Extra trees:** This is a type of tree-based ensemble method that utilizes a decision tree as the key component with a top-down method. The extra trees algorithm is considered suitable when the utilized dataset contains a significant number of continuous variables, as it decreases the computational burden and arbitrarily chooses the best feature to split (Ravi, 2020). The ET model was developed using 100 estimators.

- **(ML) Artificial Neural Networks:** Artificial Neural Networks are the most broadly utilised for predicting building energy consumption (Ahmad et al., 2014; Bourdeau et al., 2019). ANN is a non-linear computational model that emulates the functional concepts of the human brain (Amasyali and El-Gohary, 2018). ANN is an effective approach for solving non-linear problems and is dominant with big datasets which enables the neural network sufficient data to train the model (Bourhnane et al., 2020).
- **(ML) Deep Neural Network:** Deep neural network or deep learning is a machine learning technique that adopts deep patterns of the neural networks using multiple hidden layers (Hoang and Kang, 2019). The regular neural networks consist of two to three layers, limiting their capabilities to express intricate functions (Lei et al., 2021).

7.4 Data Pre-Processing

Data pre-processing has significant impacts on the machine learning algorithms performance (Kotsiantis et al., 2006). The objective of data pre-processing is to handle raw data imperfections and irregularities such as high dimensionality, noise, missing data, outliers, inconsistencies and imbalanced data (Benhar et al., 2020). This process detects the existence of invalid or inconsistent data that can cause errors during analysis (Amasyali and El-Gohary, 2018). Although data pre-processing is computationally expensive and time consuming, it is required to ensure quality assurance of the database and to avoid difficulties during model development (Shapi et al., 2021). Without the implementation of adequate pre-processing of data, several complexities could emerge that affect the model performance such as missing or abnormal values, noise etc. Missing data means the absence of one or more entries in the matrix containing the experimental dataset (Mishra et al., 2020). The missing data identified in the building and weather-related dataset were handled in different ways. Newgard and Lewis, (2015) proposed the mean value imputation for handling missing data which was utilized in the weather-related dataset. The building instances containing missing data were removed. The total number of instances removed from the building dataset is 332,150. The building dataset contains categorical features such as the roof energy efficiency and windows energy efficiency among others as shown in Figure 5.4. The categorical data were allotted values, (e.g., very good = maximum value (5) and very poor = minimum value (0)) to transform the data into an appropriate format for the ML algorithm.

7.4.1 Data Merging

The utilization of multiple datasets requires data merging. Therefore, the building dataset and meteorological dataset were merged using the common variable (postcode) to match each building data to its respective meteorological data. The datasets were merged using the Panda package of the Python programming language. The data merged resulted in a total of 15,845,526 data points.

7.4.2 Data Cleaning

The process of data cleaning involves the removal of outliers and the treatment of missing data. The meteorological data contained a few missing values which were resolved by applying the mean value imputation as proposed in 2015 by Newgard and Lewis (2015). This is the replacement of missing values with the mean value of each column. However, the instances in the building dataset with missing values (332,150 instances) were deleted from the database to avoid ambiguity and complexities during model development (training and testing phase).

7.4.3 Data Conversion

The building raw data comprised of some categorical data in variables such as wall energy efficiency, windows energy efficiency among others (see Table 5.3). The variables were allocated values [e.g., very good = highest value (5) while very poor = lowest value (0)] to provide suitable data for the ML algorithm.

7.4.4 Data Normalization

Data normalization is a very common procedure of data pre-processing that eliminates the influence of dimensions as several features often have unrelated dimensions (Y. Liu et al., 2020b; Olu-Ajayi et al., 2021). Normalization scales individual samples into a unit norm in order to avoid problems during model development. For instance, if an input variable column contains values ranging from 0 to 5, and another column holds values ranging from 1,000 to 10,000. The difference in the scale of the numbers could create problems during model development. Due to the different types of data (e.g., continuous, discrete, and categorical) present in the dataset, it is essential to normalize the data to eliminate the influence of the dimension and avoid difficulties during the model development phase (Y. Liu et al., 2020b). The building dataset and the meteorological dataset were normalized using sklearn python package normalizer. Sklearn python package is a machine learning library for the python programming language that contains various algorithm functions including pre-processing techniques such as normalization and standardization among others (Hao and Ho,

2019). Normalization utilizes the formula below to scale down the dataset such that the normalized values fall within the range of 0 and 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where: X is a data point

X_{min} minimum value

X_{max} maximum value

X_{norm} normalized value

7.5 Model Development

After completion of data pre-processing, the consolidated data was split into 70% for training and 30% which was split equally for model testing and validation. This split was performed at random based on the dependent variables to ensure sufficient classes were present in the training, test and validation data. This provides a 7:3 ratio of training to testing data. The statistical and machine learning tools utilized in this research were developed in two modes to ensure unbiased comparison. The first mode of development was conducted using various feature selection methods, while the second mode was the implementation without feature selection. Considering the availability of two dependent variables [Energy efficiency rating (categorical) and energy consumption values (Continuous)] (See section 5.5.3 for details), the classification and regression methods were employed.

In Model Development, the challenge of handling imbalanced datasets is addressed through the implementation of the Synthetic Minority Over-sampling Technique (SMOTE), effectively mitigating class imbalance by generating synthetic samples for the minority class. Within different Model Types, focus is placed on both classification and regression models. For classification tasks, statistical and machine algorithms drawn for systematic literature review (See section 3.7) such as logistic regression, decision trees, random forests, support vector machines, and artificial neural networks are employed to predict energy efficiency rating. Conversely, for regression tasks, methods like linear regression, decision trees, random forests, gradient boosting, and artificial neural networks are utilized to predict continuous numerical energy consumption values based on input features. Figure 7.1 shows the framework of this analysis.

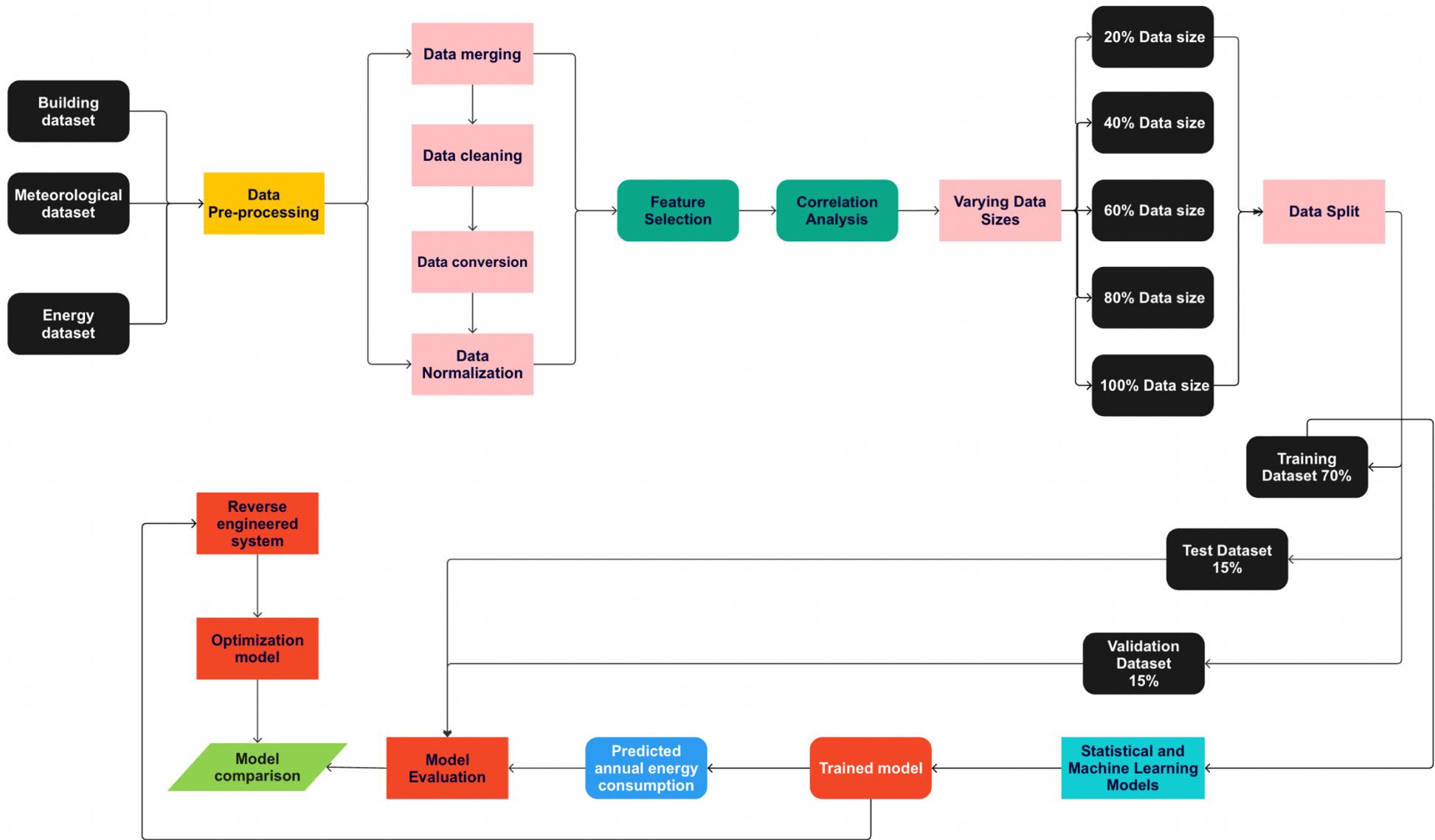


Figure 7.1: framework utilised for model development.

7.5.1 Handling Imbalanced Datasets with SMOTE

Despite the recent developments arising in machine learning and deep learning, the issue of imbalance class persists as a weighty challenge for researchers(Johnson et al., 2019) Class imbalance emerges in binary and multi-class classification tasks where selected groups are characterized as the minority class, containing fewer instances than the other groups, known as the majority class. This class imbalance issue is especially prevalent in scenarios such as medical datasets, where accurately classifying the minority class (e.g., patients with a specific disease) is vital. Typically, medical datasets are dominated by healthy patients' data, forming the majority class, while those diagnosed with the disease represent the minority class. Consequently, accurately predicting the results using an imbalanced medical dataset poses a challenge, often necessitating non-traditional ML algorithms to attain satisfactory performance. Furthermore, in imbalanced datasets, ML models essentially lean towards the majority class due to its higher occurrence, leading to the misclassification of minority class samples(Johnson et al., 2019). In the context of this research, energy consumption analysis, the majority class typically encompasses buildings with average energy ratings, as exceptional ratings (e.g., energy efficiency ratings 'A' and 'B') are finite, which makes them the minority class.

Many real-world machine learning applications which involve dealing with highly imbalanced datasets, it makes difficult for classifiers to effectively classify between minority and majority classes. Several studies have proposed the Synthetic Minority Oversampling Technique (SMOTE) as a solution to this imbalanced class problem(Xu et al., 2020; Zeng et al., 2016). Blagus and Lusa(Blagus and Lusa, 2013)conducted a comprehensive theoretical and empirical investigation into resampling techniques, particularly focusing on SMOTE. SMOTE was applied to various imbalanced datasets, real and simulated, to expound the algorithm's behaviour.

The use of Synthetic Minority Over-sampling Technique (SMOTE) to tackle imbalance in data is commonly accepted in machine learning. SMOTE is considered an effective solution to the challenges posed by imbalanced datasets. Imbalanced datasets can engender bias in machine learning models inclined towards the majority class, generating suboptimal performance, particularly for the minority class. SMOTE is a technique that generates synthetic samples for the minority class by interpolating between existing minority class instances, thus balancing the class distribution and offering the model with a more representative set of examples from all classes.

However, the decision to apply SMOTE both before and after the train-test split in a particular study specifies a thorough method to address class imbalance, although the most suitable approach remains debatable. Oversampling before the split guarantees that both testing and training sets consist of synthetic instances, aiding the model in learning and testing from augmented data, which can be particularly beneficial when the initial training set is imbalanced. Conversely, oversampling after the split is considered for a more realistic scenario, that mirrors the actual deployment of the model, allowing for evaluation of its generalization performance on real-time, non-synthetic data. Testing with data that consists of synthetic data may not precisely capture the model's performance in real-world cases. Other studies have conceptually conveyed the danger of employing oversampling before splitting data (Blagus and Lusa, 2015; Santos et al., 2018)

While SMOTE is an effective tool for handling class imbalance issues, the decision to oversample before or after the train-test split is important. In this research, oversampling was chosen due to class distribution, and it will be performed after the split to mimic real-world scenarios.

7.6.2 Overview of Classification and Regression Models

This research explores two distinctive model types to address its nine hypotheses and research objectives. These models include classification and regression. This approach allows for a comprehensive analysis of both categorical and continuous data, offering valuable insights into the research domain. These models are discussed in further detail below.

7.6.2.1 Classification Model

Classification is a supervised learning method of developing a model to forecast the class label for an unseen instance. If the quantity of class labels connected with each instance is one, it is known as Single Label Classification (SLC). SLC is grouped into binary classification and multi-class classification. Binary classification is a method of classification that includes only two different class labels such as the spam detection model (spam or not spam) and, result generation model (pass or fail). Subsequently, multi-class classification is a method of classification that includes more than two different class labels such as marital status model (married, single, divorced, widow/widower), and ethnicity detection (African American, European American, White British). This research employs the multi-class classification using the UK standard energy efficiency rating scale of A to G ('A' denoted as the most efficient and 'G' as the least efficient) as the class labels(See section 5.5.3).

Considering the type of dependent variable (categorical), the classification method was proposed which has not been explored in many previous studies for predicting energy performance. The machine learning classification method is devised for dependent variables that consist of categorical data (Loh, 2011). This method can be explained using a shape detection classification problem, as illustrated in Figure 7.2 below. The test sample (grey circle) is to be classified to predict the shape (yellow square or blue triangle). When $x = 3$ (first inner circle), the blue triangle will be selected because there are 2 triangles and just 1 square with the closest data points to the test sample. If $x = 5$ (second inner circle) the yellow square will be generated (3 squares against 2 triangles) (Olu-Ajayi, 2017). In relation to this research, the triangle and square are the seven different classes (Energy rating) which were used as the dependent variable (see Table 5.3). Furthermore, as shown in Table 5.2, the building variables are the features of the triangle (sides = 3, colour = blue) and square (sides = 4, colour = yellow). At the building design stage, the building designer will only input the required building features into the model, and the energy rating is predicted (i.e., input variables 4 and yellow, Square is predicted). To expatiate further, once an energy predictive model is developed using machine learning, when certain parameters are inputted into it, the model will predict the energy consumption level ('A', 'B', 'C'... 'G' as shown in section 5.5.3) of a building. Firstly, the model will be trained with existing data on buildings and their energy performances. After model training, the model will be tested to evaluate the accuracy as shown in section 3.5. Subsequently, after getting a satisfactory model performance, it will be deployed with a user interface which will simply request values for each variable. This deployment can be delivered in the form of a plugin or software. Furthermore, the software will be used by a building designer, to simply input values for variables related to the design such as total floor area (i.e., $20m^2$). Once these values have been inputted, the designer will run the model and receive the result in less than five minutes. The designer will receive the result, and the most important variables will also be displayed to the building designer, so the key variables can be changed if the energy performance doesn't meet the requirements or requires further improvement on energy efficiency.

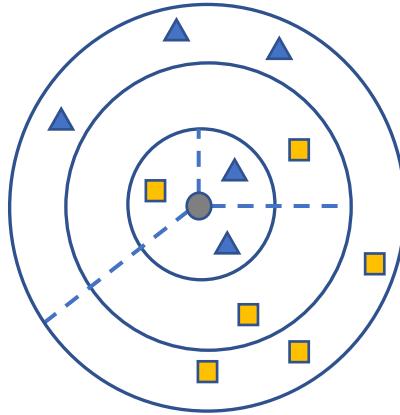


Figure 7.2: Example of ML classification method

7.6.2.2 Regression Model

Regression is also a supervised learning technique utilized in machine learning. However, unlike classification, which predicts categorical values or target class for a test instance, regression predicts continuous values (Nasteski, 2017). In the regression task, the goal was is to model the relationship between independent variables (features) and a dependent variable (target) such that given new test input data, the model can predict the corresponding output. Regression tasks are prominent in various domains such as finance, economics, and healthcare. For example, regression tasks can be utilized in predicting house prices based on features such as size, location, and number of bedrooms among others. There are various types of regression methods, including linear regression, ridge regression, lasso regression, among others. The selection of the regression method is predicated on factors such as the nature of the data and the scope of the problem being addressed. This research employs energy consumption values as the dependent variable (See section 5.5.3).

Hardware specification

The pre-processing of data, model training and testing was executed using Python programming language. These computations were performed on an Apple M1 chip MacBook Air (OS = Big Sur Version = 11.4 and with RAM = 16 GB and 8 cores). Python libraries and packages (Pandas, NumPy and Scikit-learn) were utilized for the main core of this experiment.

7.7. Model Evaluation

There are various performance measures that can be utilised for classification and regression as discussed below. For classification, the key metrics are accuracy, balanced accuracy, F1 score, and AUC while for regression the key metrics R-squared (R²), Mean Absolute Error

(MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE). These are discussed below

7.7.1 Evaluation of Classification Models

The performance of each model is evaluated using the following performance measures: Accuracy, Balanced Accuracy, F1 score, and AUC.

Accuracy: This is a type of performance measure which considered the most used in evaluating classification tasks. It is the calculation of the exact match of the estimated values and real values. Also, this measure is often used to justify that the model developed is appropriate (Gonzalez-Abril et al., 2014). The formula for accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Balanced Accuracy: Accuracy is known as the percentage of occurrences estimated correctly, while balanced accuracy is the mean of the accuracies for each class (Miller et al., 2012). There are two types of evaluation measures used to measure the balanced accuracy of the prediction result, namely sensitivity and specificity. The formula for balance accuracy is:

$$Sensitivity = \frac{TP}{TP + FN} \quad Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{Sensitivity + Specificity}{2}$$

F1 score: This is a technique for computing the weighted average of precision and recall, where a score close to one is considered the best and the score closest to zero is the worst. The formula for the F1 score is:

$$Precision = \frac{TP}{(TP + FP)} \quad || \quad Recall = \frac{TP}{(TP + FN)}$$

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

ROC AUC: This is the area under the Receiver Operator Characteristic (ROC) curve. It is largely acknowledged as one of the most appropriate pointers for classification performance. The best ROC AUC score signifying outstanding accuracy is one. However, the lowest ROC AUC score is 0.5 (Egwim et al., 2021).

7.7.2 Evaluation of Regression Models

The performance of each model is evaluated using the following performance measures: R-squared (R2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE). Among all the listed evaluation methods, the most often utilized for energy consumption prediction are the MSE and RMSE (Dong et al., 2005; Li et al., 2009b; Pham et al., 2020)

1. **Mean Absolute Error (MAE)** is a method of calculating the difference between the predicted values and the actual values at each point in a scatter plot. The closer the score is to zero, the better the performance while the higher the score, the worse the performance. It is computed as the average of the absolute errors between predicted and actual effort.

$$MAE = \frac{1}{n} \sum_{i=1}^n |AE_i - PE_i| \quad (2)$$

2. **Mean Squared Error (MSE)** is the measure of squared variation between the estimated values and the actual values. MSE is an assessment of the quality of a predictor. Models with error values closer to zero are considered the better estimation model. It is also known as Mean Squared Deviation (MSD).

$$MSE = \frac{1}{n} \sum_{i=1}^n (AE_i - PE_i)^2 \quad (3)$$

3. **Root Mean Squared Error (RMSE)** is also a metric used to calculate the differences between the estimated value and the actual perceived value of the model. It is achieved through the square root of the Mean Square Error (MSE).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (AE_i - PE_i)^2} \quad (4)$$

4. **R-squared (R²)** is a statistical measure that determines the proportion of the difference in the target variable that can be justified by the independent variables. It displays the extent to which the data fits the model. R² can produce a negative result, however, the best outcome of R² is 1.0. It is also known as the Coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{pred,i} - y_{data,i})^2}{\sum_{i=1}^n (y_{data,i} - \bar{y}_{data})^2} \quad (5)$$

7.8 Result and Discussion

In this comprehensive analysis, this section addresses nine hypotheses. Relevant features selected using the Feature selection method were employed to develop regression and classification models. For regression, a total of 12 statistical and machine learning tools were utilized, including Support Vector Regression (SVR), Artificial Neural Networks (ANN), Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), AdaBoost, Bagging, Extra Trees (ET), Ridge Regression, Lasso Regression, Bayesian Ridge (BR), Least Squares Regression (LSR), and Deep Neural Networks (DNN). Similarly, for classification, 15 statistical and machine learning tools were applied, including SVM, GB, RF, LR, KNN, DT, Naive Bayes (NB), AdaBoost, ET, Multi-layer Perceptron (MLP), Quadratic Discriminant Analysis (QDA), and Bagging. Additionally, Models were developed both with and without weather data to assess its impact. Furthermore, with a focus on reducing features, 5, 6, 7, and 10 most important features were employed to development of regression and classification models. Furthermore, reliability analysis was conducted across varying data sizes (ranging from 20% to 100%) to investigate the impact of large datasets on model performance. Additionally, efforts were made to enhance model performance through the integration of big data analytics in separate development efforts. This multifaceted approach underlines the thoroughness and rigour of the analysis, aiming to engender meaningful insights and optimize predictive modelling outcomes.

7.8.1 Feature Selection Impact on Model Performance

Firstly, considering there is no clear agreement on the effect of feature selection on the performance of the regression model (see H1 in section 1.4.2). Hypothesis H1(See section 7.8.1) was tested. Therefore, to examine the impact of the input features on regression models, the most relevant input features based on the four feature selection methods were used to develop fifteen machine learning models. However, these models will also be developed

without feature selection for a clear and unbiased comparison. Figure 7.3 shows the framework for the analysis conducted to address hypothesis H1 (See section 7.8.2), H2 (See section 7.8.3), H3 (See section 7.8.4), and H4 (See section 7.8.5).

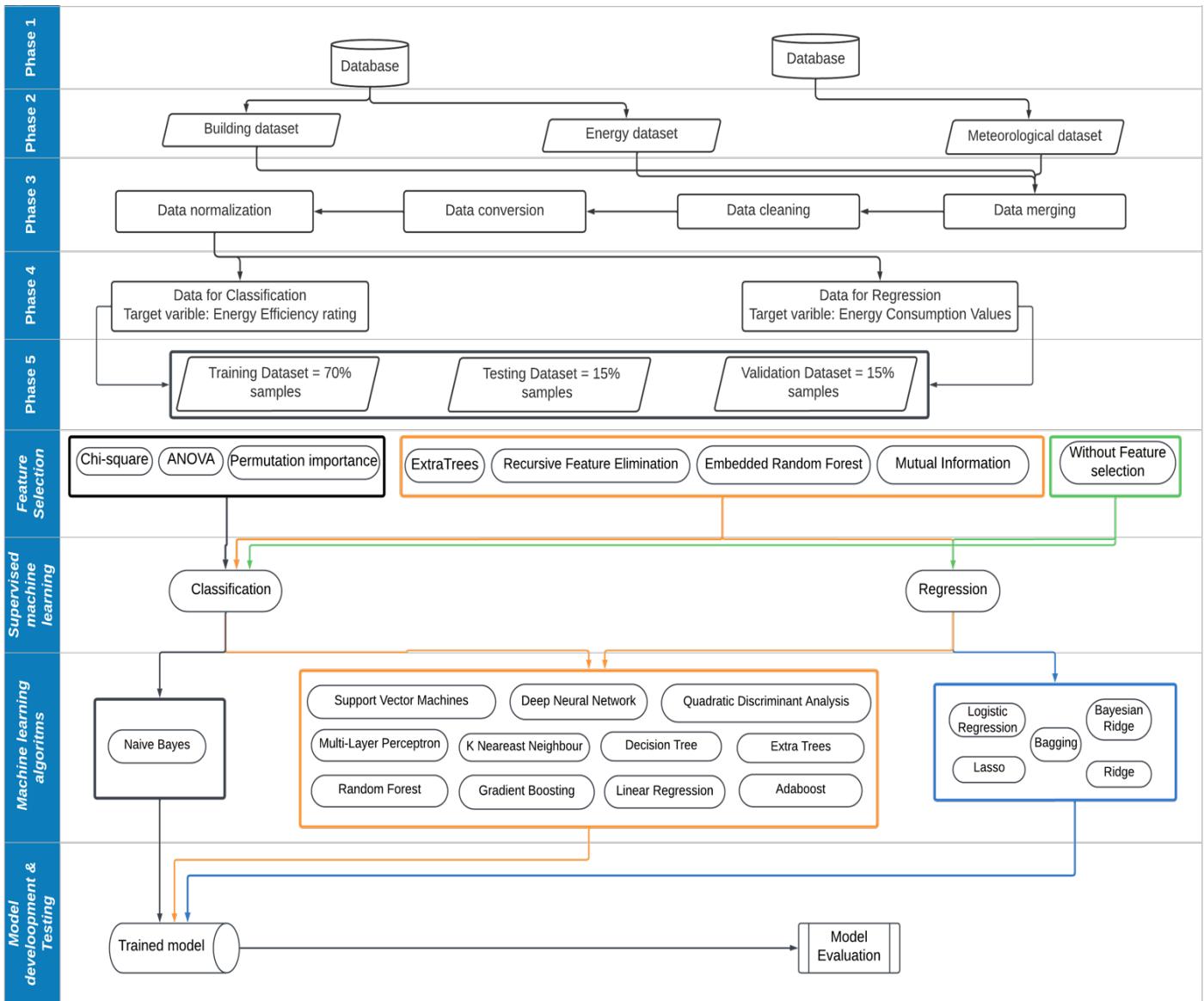


Figure 7.3: Flowchart diagram of the Feature selection impact analysis

7.8.1.1 Regression Model Performance with and without Feature Selection Methods.

In this section, this research presents the result to investigate the effect of various feature selection methods in the development of several statistical and machine learning regression models, by using four feature selection methods and fifteen statistical and machine learning for a clear comparison. This section also developed without feature selection as indicated in figure 7.3 above. Table 7.1 below show the performance results of all models with and without feature selection methods.

Table 7.1: Performance result for each model with and without feature selection methods.

	Mutual information		RFE		Random forest		Extra trees		Without FS	
MODEL	Test R-squared	Val R-squared	Test R-squared	Val R-squared	Test R-squared	Val R-squared	Test R-squared	Val R-squared	Test R-squared	Val R-squared
SVR	0.06	0.07	0.06	0.07	0.17	0.18	0.18	0.18	0.06	0.07
ANN	0.80	0.79	0.59	0.57	0.78	0.78	0.82	0.81	0.84	0.83
GB	0.85	0.85	0.78	0.78	0.84	0.84	0.84	0.83	0.86	0.86
RF	0.86	0.86	0.80	0.80	0.85	0.84	0.82	0.81	0.87	0.87
LR	0.71	0.72	0.49	0.49	0.71	0.71	0.70	0.71	0.73	0.73
KNN	0.71	0.72	0.47	0.45	0.76	0.76	0.75	0.75	0.64	0.63
DT	0.73	0.73	0.63	0.65	0.71	0.71	0.71	0.69	0.75	0.73
ADABOOST	0.68	0.67	0.42	0.42	0.64	0.64	0.64	0.63	0.68	0.68
BAGGING	0.84	0.84	0.78	0.79	0.83	0.83	0.81	0.80	0.85	0.86
ET	0.85	0.84	0.79	0.80	0.84	0.83	0.79	0.78	0.87	0.86
RIDGE	0.71	0.72	0.49	0.49	0.71	0.71	0.70	0.71	0.73	0.73
LASSO	0.71	0.72	0.49	0.49	0.71	0.71	0.70	0.71	0.73	0.73
BR	0.71	0.72	0.49	0.49	0.71	0.71	0.70	0.71	0.73	0.73
LSR	0.04	0.02	-0.06	-0.08	0.24	0.27	-0.14	-0.13	0.01	-0.03
DNN	0.82	0.81	0.70	0.69	0.84	0.83	0.83	0.82	0.84	0.83

7.8.1.2 Classification Model Performance with and without Feature Selection Methods.

In this section, this research presents the result to investigate the effect of various feature selection methods in the development of several machine learning regression models, by using seven feature selection methods and twelve statistical and machine learning tool for a clear comparison. Similar to the regression tasks, this section also developed without feature selection as indicated in figure 7.3 above. Table 7.1 below show the performance results of all models with and without feature selection methods.

Table 7.1: Performance result for each model with and without feature selection methods.

MODEL	Chi-square		Anova		Mutual Information		Permutation importance		RFE		Random forest		Extra Trees		Without FS	
	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val
SVM	0.59	0.61	0.60	0.61	0.59	0.60	0.56	0.57	0.46	0.48	0.57	0.58	0.57	0.57	0.43	0.46
GB	0.69	0.69	0.69	0.69	0.69	0.69	0.65	0.66	0.58	0.59	0.69	0.69	0.68	0.69	0.67	0.68
RF	0.66	0.66	0.66	0.67	0.67	0.67	0.61	0.61	0.57	0.58	0.69	0.69	0.69	0.69	0.66	0.65
LR	0.58	0.58	0.58	0.59	0.57	0.58	0.50	0.52	0.47	0.47	0.54	0.54	0.56	0.55	0.56	0.53
KNN	0.59	0.60	0.62	0.64	0.63	0.64	0.59	0.60	0.48	0.49	0.62	0.62	0.61	0.62	0.58	0.58
DT	0.59	0.60	0.63	0.62	0.60	0.60	0.60	0.60	0.48	0.50	0.60	0.61	0.60	0.60	0.58	0.58
NB	0.54	0.54	0.55	0.54	0.56	0.55	0.50	0.50	0.45	0.44	0.52	0.52	0.54	0.54	0.21	0.20
ADABOOST	0.54	0.53	0.51	0.50	0.58	0.58	0.61	0.62	0.53	0.55	0.59	0.60	0.51	0.51	0.53	0.54
ET	0.64	0.65	0.64	0.66	0.65	0.65	0.61	0.61	0.55	0.56	0.68	0.68	0.68	0.67	0.64	0.63
MLP	0.67	0.68	0.67	0.68	0.68	0.69	0.61	0.63	0.57	0.58	0.68	0.69	0.68	0.68	0.55	0.53
QDA	0.47	0.46	0.29	0.27	0.13	0.12	0.52	0.53	0.36	0.35	0.50	0.50	0.47	0.47	0.53	0.56
BAGGING	0.64	0.65	0.64	0.65	0.65	0.66	0.61	0.61	0.54	0.55	0.66	0.66	0.67	0.66	0.65	0.63

7.8.2 H1: Feature Selection does not influence Regression Model Performance.

Some studies suggest that feature selection is essential for the optimum performance of the model (Alaka et al., 2018; Balogun et al., 2021; Zhang and Wen, 2019a; Hai-xiang Zhao and Magoulès, 2012a); some studies have argued that feature selection is more effective in classification than regression prediction (Jović et al., 2015; Kumar, 2014). (see section 1.4.2 for more justification). Thus, it is hypothesised that:

- **Null hypothesis H_0 :** Feature selection does not influence regression model performance.
- **Alternative hypothesis H_A :** Feature selection has an effect on regression model performance.

The basis of determining the best regression predictive model stated in section 7.7.2 stipulates that models holding values closer to zero for MAE, MSE and RMSE are the good predictive models while values closer to one for r2 produced the best results. For the purpose of the comparison of model performance with and without feature selection, r2 was employed. Table 7.2 shows the result from the test and validation set to eliminate the possibility of overfitting. In this research, GB and ET produced the best result comparatively for predicting annual energy consumption. Although GB achieved the highest r2 value of 0.87, this was achieved

without feature selection. While it is noted that feature selection improves the performance of ML models by eliminating the unimportant and irrelevant noisy features, thus improving the quality of the dataset (Asir et al., 2016), Ten statistical and machine learning models produced better performance without FS than with the application of feature selection, namely GB, RF, ANN, LR, Bagging, ET, BR, ridge, Lasso and BR. For example, ANN achieved an r^2 score of 0.83 without feature selection and produced 0.57 using the RFE feature selection method. However, some models did produce better performance with than without the application of feature selection such as SVM and KNN. While few algorithms showed no notable improvement in performance with and without specific feature selection such as DNN and DT. For example, DNN produced the same r^2 performance value of 0.83 with and without the application of RF feature selection. Whereas the r^2 outcome of SVR using RF is 0.18 and 0.7 without feature selection. It is concluded that feature selection can have favourable and unfavourable impacts on statistical and machine learning algorithms depending on the type of algorithm employed in a regression task.

The bar plot (Figure 7.4) derived from Table 7.2, displays the model's prediction performance of fifteen statistical and machine learning tools with and without the application of different feature selection methods. It provides clear evidence to reject the alternative hypothesis, denoting that feature selection does not indeed have an effect on regression model performance. The application of feature selection methods, such as Mutual Information, Recursive Feature Elimination (RFE), Random Forest, and Extra Trees did not generally lead to improved performance values, as shown by higher R-squared values in Figure 7.4 below. These results align with the hypothesis suggesting that the application of feature selection does not necessarily have significant effect on the performance of regression models(see section 1.4.2). However, it's essential to recognise that the impact of feature selection is not unanimous across all algorithms. Two algorithms namely KNN and SVM showed significant increase in performance using feature selection. KNN shows significant increase in performance using RF with an r^2 value of 0.76 and 0.63 without the application of feature selection. Despite these exceptions, the majority of the models show limited positive and negative impacts on performance using feature selection, emphasizing the impractical importance of feature selection in optimizing regression models. Therefore, while H_A may hold true for specific algorithms, the predominant trend supports the deduction that feature selection does not have a notable positive impact on regression model outcomes. Therefore, the alternative hypothesis (H_A) is rejected. However, the alternative hypothesis can hold true in studies dependent on the algorithm utilised. For example, Kapetanakis et al., (2017) employed ANN to develop regression models for predicting energy consumption with and without feature selection. It was

established that the performance with the feature selection did not outperform the performance without the application of feature selection. Similarly, in this investigation, ANN is one of the distinct models that do not elicit better performance with the application of feature selection. Nevertheless, the variation observed across various algorithms emphasize the need for a tailored approach and more informed selection of feature selection methods, as their efficacy may vary depending on the essential characteristics such as the type of algorithm selected.

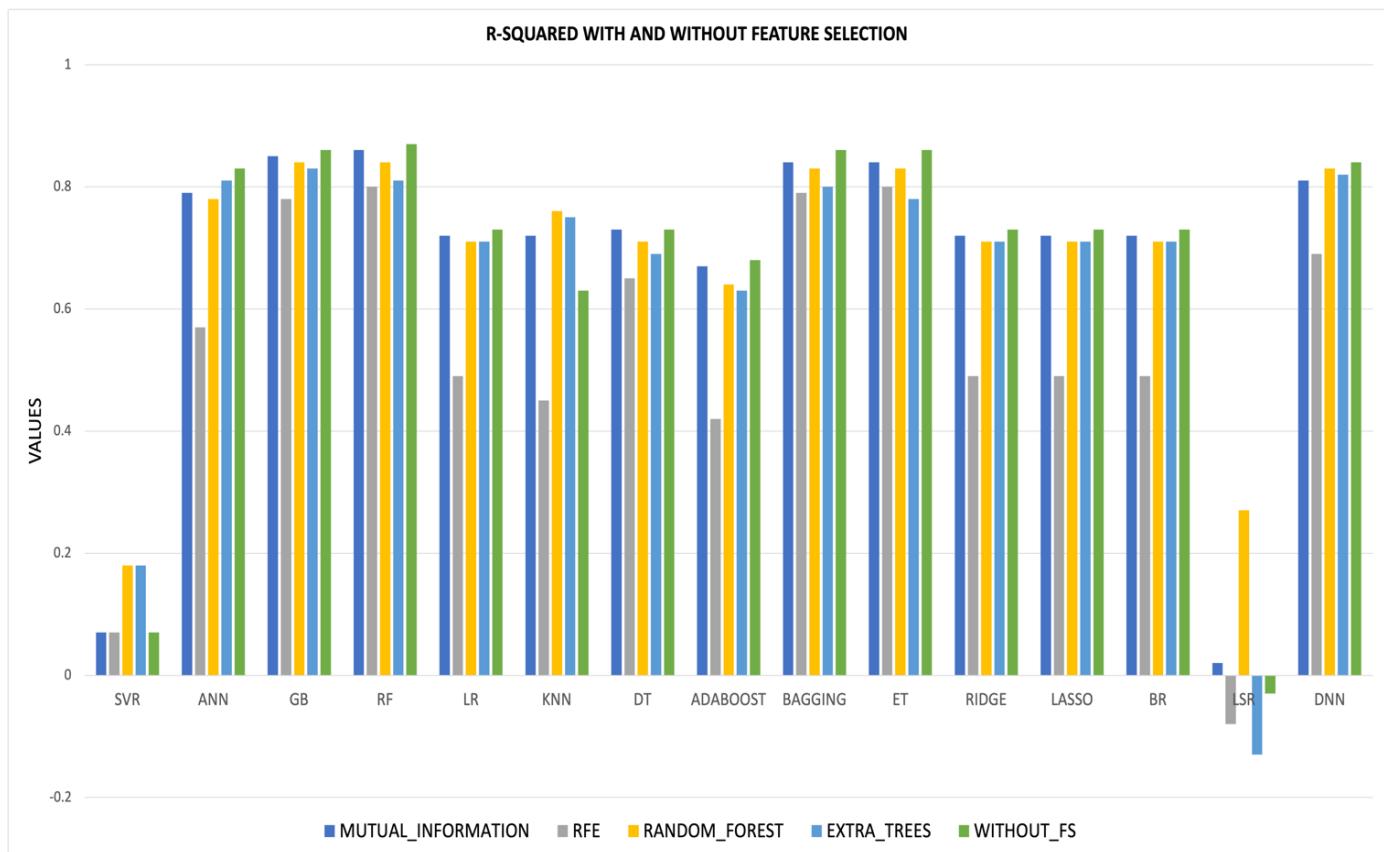


Figure 7.4: Prediction performance distribution for each ML algorithm with and without FS methods(Regression)

It is well noted that this result does not demonstrate that feature selection is more effective in classification than regression prediction. Therefore, for a clear and unbiased comparison of the efficacy of feature selection, this research employs seven feature selection methods in the development of twelve statistical and machine learning classification models.

For classification models, the performance measures for selecting the most effective predictive model are accuracy, balanced accuracy, and F1 score among others, as outlined in section 7.7.1. Table 7.2 and Figure 7.5 visualizes a clear comparison of the negative and positive impacts the FS selection methods can have on the predictive accuracy of the twelve statistical and ML models. It is evident that similarly, feature selection can have a positive or negative impact on

the predictive accuracy, depending on the algorithm selected. However, all models produced a better performance with the application of feature selection, except one namely QDA. Gradient Boosting (GB), Random Forest (RF), and Extra Trees (ET) consistently produce good accuracy, showing stable performance values with minimal difference between the test and validation values. In contrast, Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT) among others produced good accuracy with sensitivity to feature selection, resulting in notable differences in accuracy across methods. Models such as Naive Bayes (NB), AdaBoost, Bagging, and Multi-layer Perceptron (MLP) displayed varied performance, indicating sensitivity to feature selection with notable differences in accuracy across methods. However, Quadratic Discriminant Analysis (QDA), on the other hand, consistently produced poor results across all feature selection methods. GB, RF, and MLP are top-performing models based on accuracy. The consistent poor performance of QDA suggests limitations in the algorithm's ability to handle the small features, as corroborated by (Wang et al., 2008). Wang et al., (2018) conducted a comparison between Linear discriminant analysis and QDA in a small size and it was noted that LDA produced better performance. The Gradient Boosting (GB) model emerged the best with and without feature selection, it achieved a better accuracy of 0.70 with MI and PI feature selection method as compared to 0.65 without feature selection. Contrarily, Quadratic Discriminant Analysis (QDA) generated better predictive accuracy of 0.56 without FS and 0.12 with FS.

Based on result in Figure 7.3 and 7.5, it can be concluded that feature selection is more effective for prediction in classification tasks than regression tasks. Consequently, regardless of the task, feature selection can have negative or positive impact on model performance. This is in line with previous research, which states that the achievement of a good predictive accuracy of a model is highly predicated on the type of algorithm and feature selection method chosen (Balogun et al., 2021; Olu-Ajayi et al., 2022b). Therefore, if an appropriate feature selection method is not selected for the specific ML algorithm used, feature selection will not result in good accuracy. For example, SVM produced good results for Chi-square, Mutual Information and ANOVA, which are all filter methods but performed poorly using RFE.

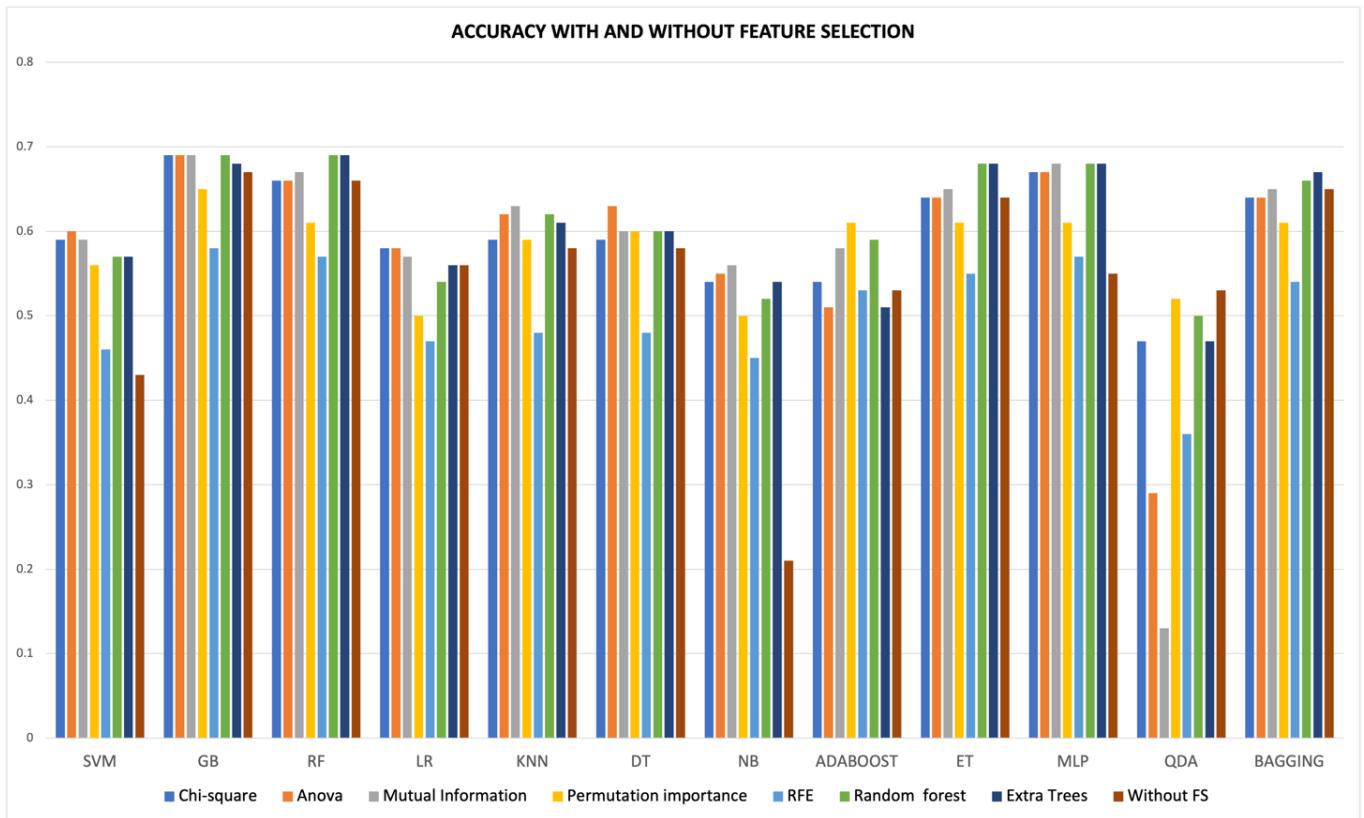


Figure 7.5: Prediction performance distribution for each ML algorithm with and without FS methods (Classification)

Previous studies have applied the feature selection method in the development of energy predictive models (M. W. Ahmad et al., 2017; Z. Dong et al., 2021; Zhang & Wen, 2019b). Nonetheless, the fraction of studies that deliver comprehensive insights into the incorporation and capabilities of feature selection with machine learning is still limited(Olu-Ajayi et al., 2023b). The results of the analysis conducted (as displayed in Figure 7.4 and Figure 6.9) show that feature selection can have positive or negative impacts on the statistical and machine learning model dependent on the feature selection method and algorithms. Therefore, based on the result achieved the chart below shows the feature selection methods that have positive impacts on various model performances.

Table 7.2: feature selection suitability for certain algorithms

Type	Feature Selection method	Regression	Classification
FILTERS	Chi-Square	None	All algorithms except NB
	Mutual Information	SVR, KNN	All algorithms except NB
	Anova	None	All algorithms except NB
	Permutation Importance	None	All algorithms except NB
WRAPPER	Recursive Feature Elimination (RFE)	None	NB, ADABOOST
EMBEDDED	Embedded Random Forest (ERF)	SVR, RF, KNN, DT	All algorithms except NB
	Extra trees (ET)	SVR, KNN, DT, ADABOOST,	All algorithms except NB

7.8.3 H2: Feature Selection has Positive Impacts on Machine Learning Energy Prediction Model Performance.

Feature selection has been noted for its importance in the achievement of high machine learning prediction according to the majority of studies (e.g., Ahmad et al., 2017; Dong et al., 2021a; Zhang and Wen, 2019 among many others). It has been argued by few studies that carried out experimental research that feature selection can have negative impacts on ML model performance (Balogun et al., 2021; Kapetanakis et al., 2017). (see section 1.4.2 for more justification). Thus, it is hypothesised that:

- **Null hypothesis H_0 :** Feature selection has positive impacts on machine learning energy prediction model performance.
- **Alternative hypothesis H_A :** Feature selection does not have a positive impact on machine learning energy prediction model performance.

The result from the classification task and regression task displayed in Table 7.2 and Table 7.3 respectively indicates that feature selection does have a positive impact on machine learning model performance, which supports the acceptance of Null hypothesis H_0 . In the classification task, the application of feature selection methods such as Chi-square, ANOVA, MI, PI, among others led to improved performance values based on accuracy for various algorithms such as SVM, GB, RF, KNN, among others. The positive impacts of feature selection methods are apparent in the accuracy of these models in comparison to the accuracy of models developed without feature selection as shown in Figure 7.5. Generally, at least one feature selection method had positive impacts on 9 out of 10 machine learning models. While for regression tasks, some algorithms such as SVR, KNN, and AdaBoost produced better test and validation R-squared values when feature selection methods like MI, RFE, and ET were employed. Although the result has a negative impact on specific algorithms, the result indicates that selecting the most relevant features contributes to improved performance.

In the regression task, the application of feature selection methods, specifically Mutual Information (MI) Random Forest (RF) and Extra trees(ET), exhibited positive impacts on the performance of several models, namely KNN, DNN, and SVR. This enhancement was detected in 3 out of the total 10 machine learning algorithms considered. However, it is worth noting that statistical feature selection methods, despite minute, showed improvements in SVR, indicating that different algorithms respond differently to feature selection techniques. Further

reinforcing the hypothesis that feature selection does have positive effect on regression model performance. Therefore, the null hypothesis(H_0) is accepted.

7.8.4 H3: ML Feature Selection Methods lead to Better Performance of Machine Learning Prediction Models than Statistical Feature Selection Methods.

In the classification task, machine learning feature selection methods (MLFS) (RF & ET) and statistical feature selection(SFS) methods showed leading to improvements in performance across all 10 machine learning algorithms. However, it should be noted that statistical feature selection methods showed limited improvements in Bagging and Random Forest compared to other machine learning models. To answer the following hypotheses:

- **Null hypothesis H_0 :** ML feature selection methods lead to better performance of machine learning prediction model than statistical feature selection methods.
- **Alternative hypothesis H_A :** ML feature selection methods do not lead to better performance of machine learning prediction models than statistical feature selection methods.

The observed improvements in both regression and classification tasks (see in Table 7.2 and Table 7.3) supports the null hypothesis, suggesting that machine learning feature selection methods indeed contribute to better overall model performance compared to statistical feature selection methods. This is in alignment with other studies such as (Guo et al., 2020a; Nguyen et al., 2013). For example, Guo et al., (2020) applied Boruta feature selection in the development of energy prediction models using random forest(RF), and SVM and produced good accuracy of 82% and 90% respectively. Therefore, the null hypothesis(H_0) is accepted. However, the instances where statistical methods exhibited improvement in performance are specific to certain algorithms, emphasizing the need for an informed technique of selecting feature selection methods based on the characteristics of the task and the primary machine learning model. Using the example of SVR, only the machine learning feature selection method engendered performance. Given these results, it is also worth noting that some studies have argued that they do not perform best when using the same algorithm for feature selection and model development. For example, Ahmad et al., 2017a utilized random forest feature selection in the development of Random Forest (RF) and Artificial Neural Networks (ANN) building energy prediction model.

7.8.5 H4: Using the Same Algorithm for Feature Selection and Prediction leads to better Model Performance than Using Different Algorithms.

Some experimental studies contend that the application of random forest for feature selection and model development can significantly improve model performance (Huljanah et al., 2019; Nguyen et al., 2013). For examples, Nguyen et al., (2013) random forest for feature selection and model development in predicting prostate cancer and achieved 99% accuracy. (see section 1.4.2 for more justification). Thus, it is hypothesised that:

- **Null hypothesis H_0 :** Using the same algorithm for feature selection and prediction leads to better model performance than using different algorithms for both tasks.
- **Alternative hypothesis H_A :** Using the same algorithm for selection and prediction does not necessarily lead to better model performance than using different algorithms for both tasks.

Although MLFS methods have been proffered to be most suitable for better ML model performance in many studies (M. W. Ahmad et al., 2017; Z. Dong et al., 2021; Zhang & Wen, 2019b), some studies have argued that they do not perform best when using the same algorithm for feature selection and model development. For example, Ahmad et al., 2017a utilized random forest feature selection in the development of Random Forest (RF) and Artificial Neural Networks (ANN) building energy prediction model. Figure 7.6a-b(derived from result Table 7.2 – 7.3) shows the result of RF and ET across all feature selection methods in classification and regression tasks for a clear and unbiased comparison. For the classification task, the result indicates that machine learning feature selection methods (RF & ET) do lead to an increase in performance of the Random Forest (RF) and Extra Trees (ET) models, and the difference appears to be relatively significant in comparison to other feature selection methods such as Chi-square and ANOVA. Therefore, this supports the acceptance of the null hypothesis (H_0): Using the same algorithm for feature selection and prediction leads to better model performance than using different algorithms for both tasks. However, in the regression context, the utilization of machine learning feature selection methods (RF & ET) did demonstrate good performance for RF and ET models in comparison to other FS methods except MI FS method.

Therefore, based on the results, the null hypothesis is accepted, denoting that employing the same algorithm for both feature selection and prediction can be beneficial in certain situations, particularly in classification tasks and essentially based on the type of data utilised.

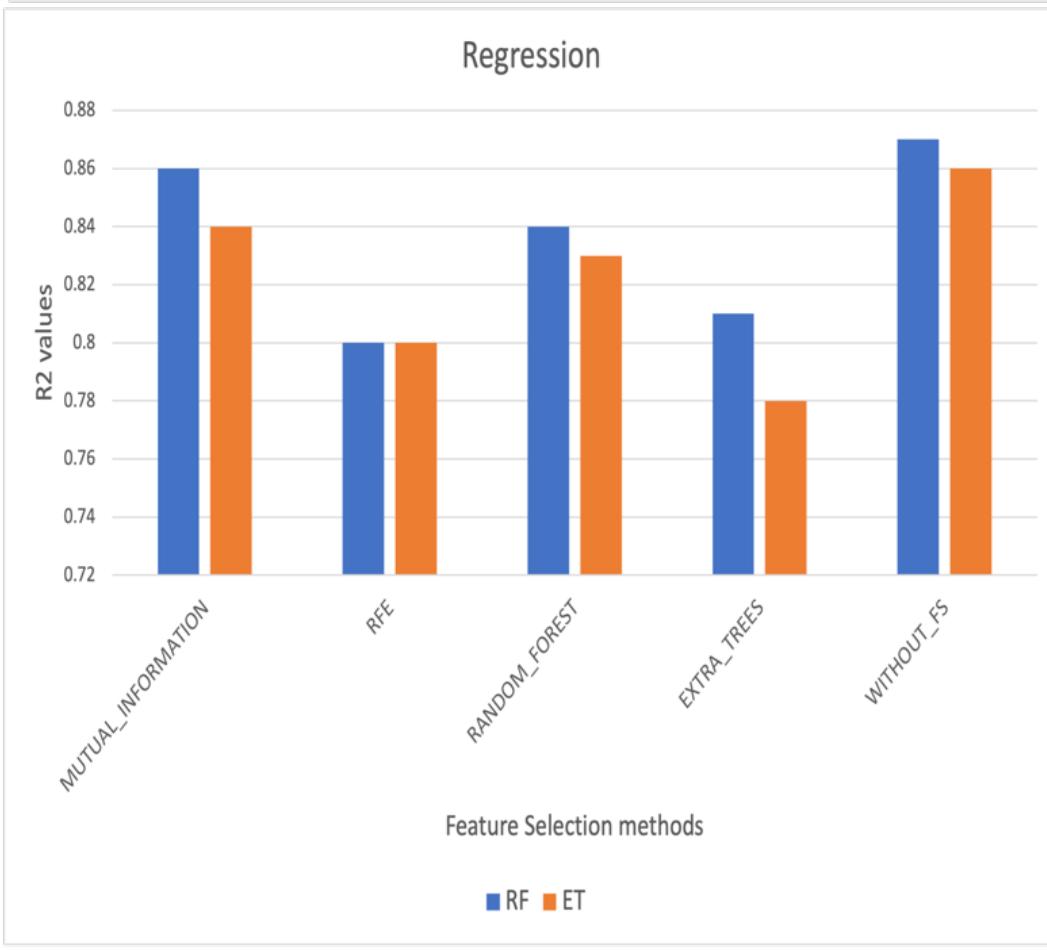
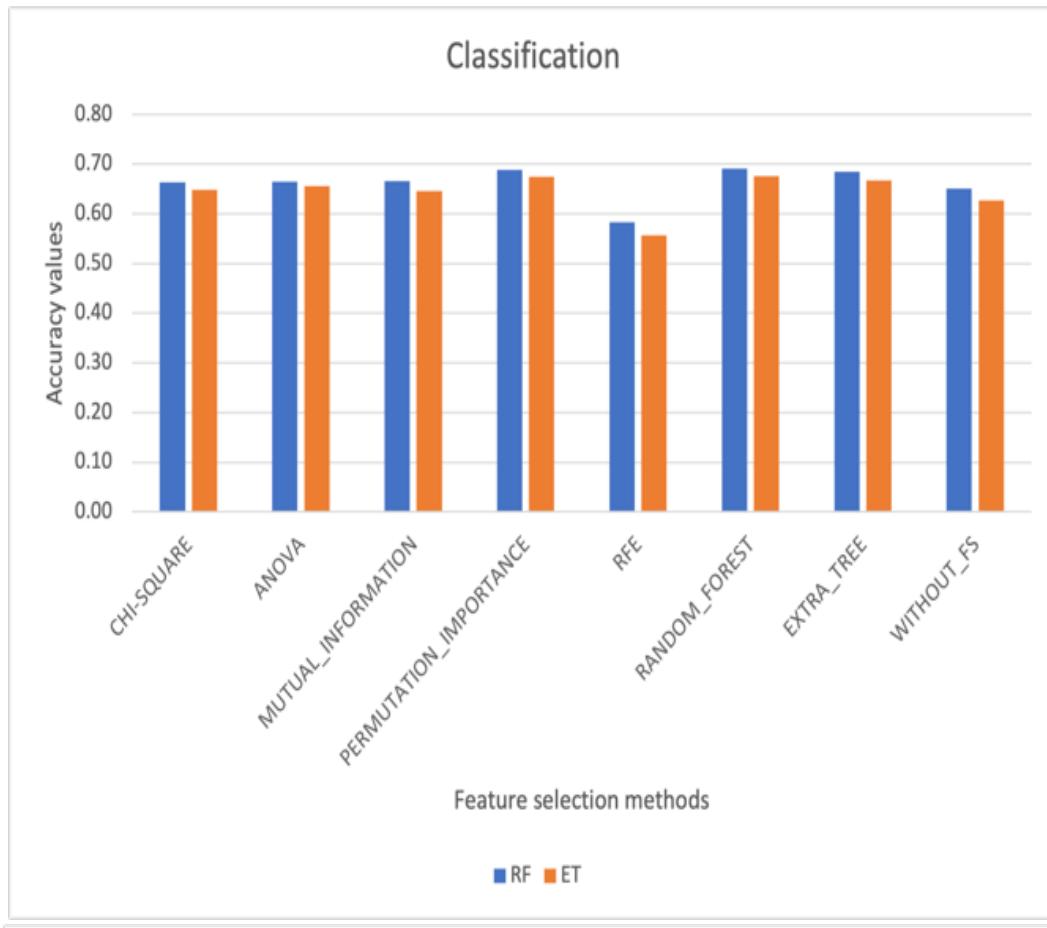


Figure 7.6: Prediction performance distribution for ML feature selection methods (classification & regression)

7.8.6 H5: Weather Data Does not Significantly Improve ML Model Performance for Building Energy Consumption Prediction.

In this research, meteorological data was employed based on the proffered importance for energy consumption prediction and multiple research studies have employed weather or meteorological data in the development of machine learning models for predicting energy consumption, and they have conveyed good model performance (for instance, Bagnasco et al., 2015; Ding and Liu, 2020; Dong et al., 2005, 2021b, p. 2; Kim et al., 2020, and others). To illustrate, Kamel et al. (2020) used weather data and employed eXtreme Gradient Boosting (XGB) algorithm to predict building energy consumption, achieving a Root Mean Square Error (RMSE) of 0.038. However, it is noteworthy that certain studies, which did not utilize weather data when developing machine learning models for energy consumption prediction, achieved comparable results. For instance, Almalaq and Zhang (2019) attained an RMSE of 0.186 using ANN for energy consumption prediction without including weather data. Similarly, Izidio et al. (2021) accomplished an RMSE of 0.077 using Support Vector Machines (SVM) for building energy consumption prediction. Consequently, it is challenging to assert that weather data significantly influences the performance of machine learning models. Moreover, none of the studies conducted a comprehensive and unbiased comparison of model performance with and without the inclusion of weather data to validate the impact of weather data on model performance. The following hypothesis was formulated:

- **Null hypothesis H_0 :** Weather data do not significantly improve ML model performance for building energy consumption prediction.
- **Alternative hypothesis H_A :** Weather data significantly improves ML model performance for building energy consumption prediction.

Based on the observed result shown in Figures 6.2 and 6.4, it was noted that some feature selection methods such as mutual information and random forest did not capture weather features as very relevant, while some did not select weather features as relevant such as chi-square, ANOVA and permutation importance. The Pearson correlation analysis (see section 6.5.1.2 and 6.5.2.2) does not show strong correlation between weather variables and energy efficiency rating and energy consumption values. However, it is noted that this is not enough to conclusively validate the importance of meteorological data.

Therefore, adopting phase 1 – 5 in the figure 7.3, all tools were employed to develop models with and without meteorological dataset to further asses the effect of weather data on model

performance. Figure 7.7 and Figure 7.8 show the model performance with and without weather variables in regression and classification tasks respectively.

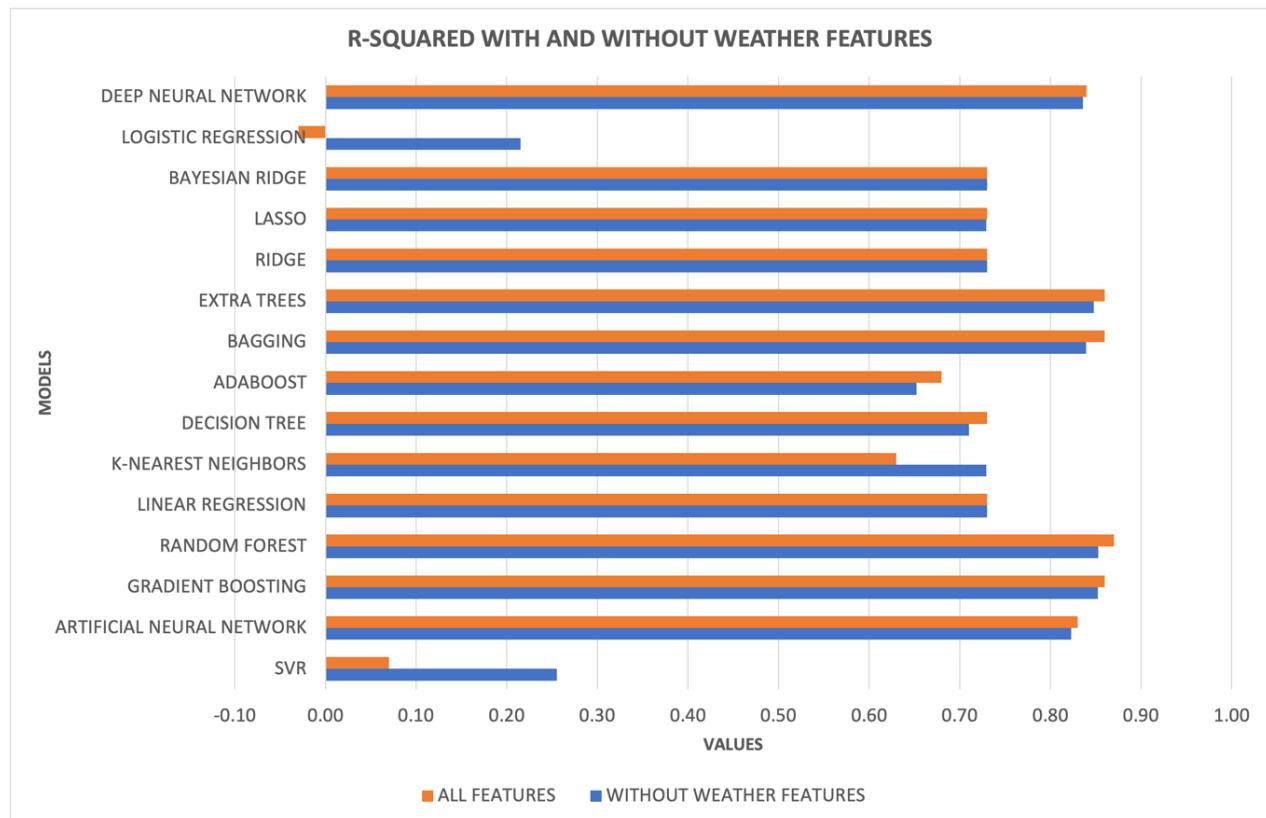


Figure 7.7: model's performance(r^2) with and without weather data (regression)

Firstly, In the context of model performance, significant improvement refers to a notable enhancement in the model's predictive accuracy. For instance, in a classification model, a significant improvement means increasing the accuracy from 85% to 95%, reducing the error rate by 10 percentage points. In regression models, it could involve a increase in the r^2 from 0.10 to 0.20, indicating a more precise prediction. The result generated shows that the addition of the weather data elicits better performance across all models except LSR, and KNN, for regression task as shown in Figure 7.7 above. However, some models produced the same performance with and without weather features such as BR, RIDGE, LASSO and linear regression, The support vector machine model showed considerably better performance without the feature selection method in line with its prominence based on its ability to produce the optimum results using small sample sizes (Aversa et al., 2016; Mat Daut et al., 2017; Olu-Ajayi et al., 2021). Given that only a few feature selection methods namely MI, RFE and RF selected weather features and realistically, weather data does influence energy consumption. Therefore, it can be concluded that not all feature selection methods perform well in their

unitary state without domain knowledge. Nevertheless, although the addition of weather data does improve the performance, it does not engender a significant increase. Furthermore, in the classification task, Figure 7.8 shows that the addition of the weather data does not engender better accuracy values across models except Quadratic Discriminant Analysis(QDA). This can be subject to the inability of QDA to produce good performance in a small sample size.(Wang et al., 2008) Based on the observed result, weather data does have minimal effect on model performance, therefore the null hypothesis is accepted which stipulates that Weather data do not significantly improve ML model performance for building energy consumption prediction.

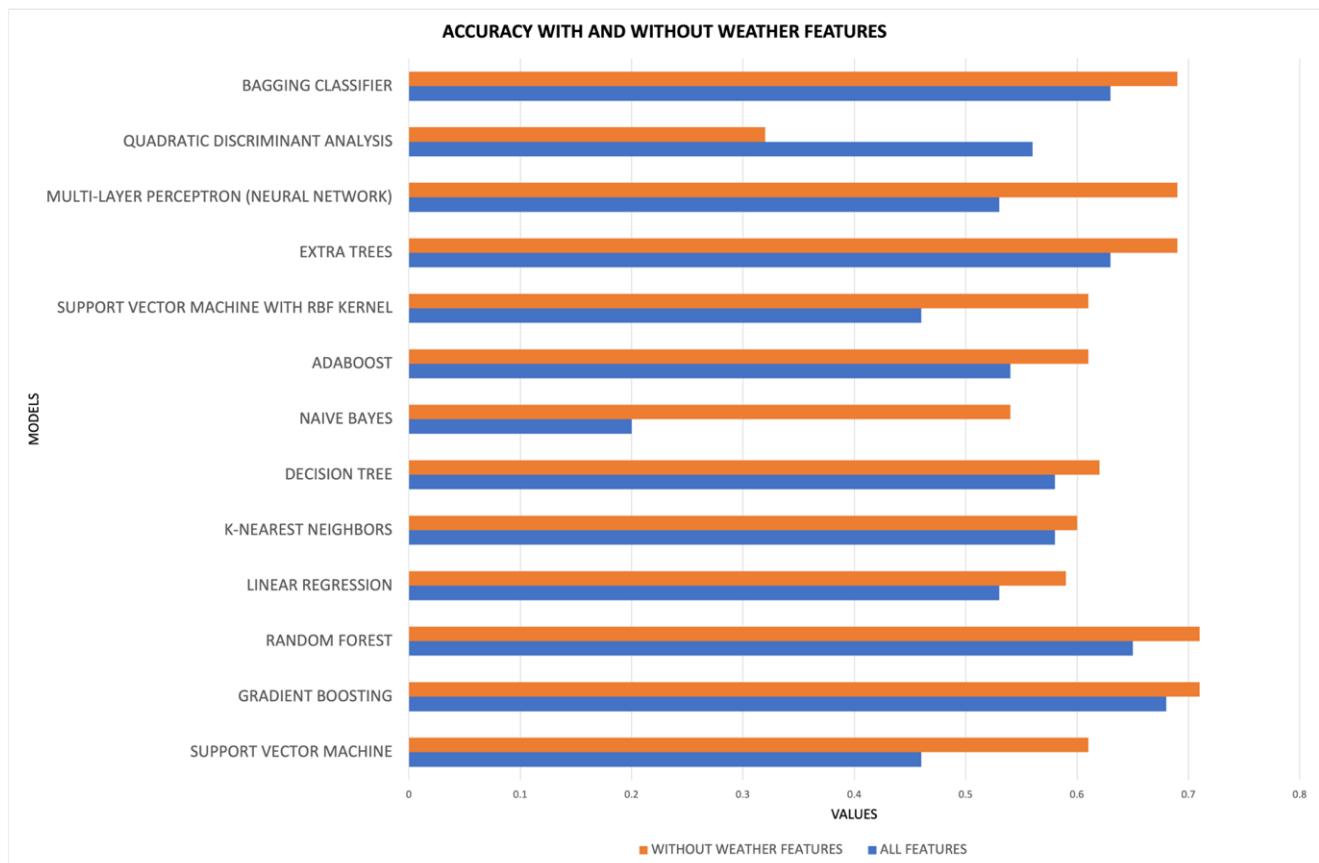


Figure 7.8: model's performance(accuracy) with and without weather data

7.8.7 Key Feature Impact Analysis

Furthermore, considering the primary objective is to streamline the number of features for energy assessment at the design stage, by identifying and employing a reduced set of relevant features in the model development and limiting the number of features required by building designers. Albeit, various feature selection methods were employed to achieve this, it resulted in over 11 selected features with the exception of permutation importance which produces suboptimal performance. Recognizing the impact of practicality and efficiency, an experimental approach was conducted to assess whether a subgroup of the most important features, specifically five, six, and seven features, could produce comparable or exact

performance values to models developed utilizing the originally identified set of over 11 features.

The experiment is essential a systematic assessment of algorithms' predictive capabilities using 5, 6, 7 and 10 of the most important features deduced by the different feature selection methods. As shown in the framework in Figure 7.3, the phases 1 to 5 remain the same. However, the various models were developed with 5, 6, 7 and 10 of the most important features of each feature selection method for classification and regression tasks. This iterative procedure enables a detailed exploration of model complexity and performance, with the goal of finding an optimal balance. This approach is also paramount for practical applications, as it allows for the identification of a subset of features that not only contributes significantly to the model's performance but also it can also facilitate a more streamlined and resource-efficient execution in real-world scenarios. This experiment is essential a systematic assessment of algorithms' predictive capabilities using 5, 6, 7 and 10 of the most important features

7.8.7.1 Impact on Classification Model Performance

In a classification task, this section developed statistical and machine learning model for 5,6 and 7 of the most important features for each feature selection method namely Chi-Square, Anova, Extra Trees, Mutual Information, Permutation Importance, Random Forest and Recursive Feature Elimination. Table 7.3 shows the performance values of the model using

5 FEATURES	CHI-SQUARE				ANOVA				EXTRA TREES				MUTUAL INFORMATION				PERMUTATION IMPORTANCE				RANDOM FOREST				RECURSIVE FEATURE ELIMINATION			
MODEL	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SVM	0.53	0.51	0.53	0.50	0.52	0.51	0.52	0.49	0.56	0.55	0.56	0.55	0.52	0.51	0.52	0.49	0.56	0.52	0.56	0.54	0.50	0.45	0.50	0.46	0.45	0.20	0.45	0.28
GB	0.63	0.63	0.63	0.61	0.62	0.64	0.62	0.61	0.63	0.62	0.63	0.61	0.63	0.64	0.63	0.61	0.65	0.64	0.65	0.64	0.62	0.60	0.62	0.60	0.51	0.49	0.51	0.47
RF	0.60	0.60	0.60	0.60	0.62	0.63	0.62	0.61	0.63	0.62	0.63	0.61	0.63	0.63	0.63	0.61	0.62	0.61	0.62	0.61	0.59	0.58	0.59	0.58	0.47	0.45	0.47	0.46
LR	0.44	0.40	0.44	0.41	0.56	0.53	0.56	0.54	0.52	0.48	0.52	0.48	0.54	0.50	0.54	0.51	0.50	0.45	0.50	0.43	0.49	0.45	0.49	0.45	0.44	0.43	0.44	0.30
KNN	0.57	0.56	0.57	0.56	0.59	0.59	0.59	0.57	0.58	0.56	0.58	0.57	0.60	0.59	0.60	0.58	0.59	0.58	0.59	0.56	0.55	0.56	0.55	0.46	0.43	0.46	0.44	
DT	0.60	0.59	0.60	0.59	0.62	0.63	0.62	0.61	0.63	0.61	0.63	0.61	0.63	0.63	0.63	0.61	0.60	0.59	0.60	0.59	0.59	0.58	0.59	0.47	0.44	0.47	0.45	
NB	0.49	0.51	0.49	0.46	0.50	0.51	0.50	0.48	0.50	0.50	0.50	0.48	0.51	0.51	0.51	0.48	0.50	0.48	0.50	0.49	0.49	0.49	0.49	0.44	0.42	0.44	0.39	
ADABOOST	0.49	0.51	0.49	0.47	0.61	0.61	0.61	0.59	0.62	0.60	0.62	0.61	0.60	0.60	0.60	0.58	0.61	0.60	0.61	0.59	0.60	0.59	0.60	0.50	0.45	0.50	0.45	
ET	0.60	0.59	0.60	0.59	0.62	0.63	0.62	0.61	0.63	0.61	0.63	0.61	0.63	0.63	0.61	0.61	0.60	0.61	0.60	0.59	0.58	0.59	0.58	0.47	0.45	0.47	0.45	
MLP	0.61	0.59	0.61	0.58	0.62	0.63	0.62	0.60	0.62	0.61	0.62	0.61	0.62	0.62	0.60	0.61	0.60	0.61	0.60	0.60	0.58	0.60	0.57	0.50	0.47	0.50	0.46	
QDA	0.45	0.52	0.45	0.43	0.18	0.33	0.18	0.19	0.52	0.53	0.52	0.50	0.49	0.53	0.49	0.50	0.52	0.50	0.52	0.51	0.50	0.49	0.50	0.49	0.06	0.37	0.06	0.09
BAGGING	0.60	0.59	0.60	0.59	0.62	0.63	0.62	0.61	0.63	0.62	0.63	0.61	0.63	0.63	0.61	0.61	0.60	0.61	0.61	0.59	0.58	0.59	0.58	0.47	0.44	0.47	0.45	

6 FEATURES	CHI-SQUARE				ANOVA				EXTRA TREES				MUTUAL_INFORMATION				PERMUTATION IMPORTANCE				RANDOM FOREST				RECURSIVE FEATURE ELIMINATION			
MODEL	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SVM	0.53	0.51	0.53	0.50	0.57	0.54	0.57	0.55	0.56	0.53	0.56	0.54	0.53	0.52	0.53	0.50	0.56	0.52	0.56	0.53	0.50	0.45	0.50	0.46	0.44	0.29	0.44	0.28
GB	0.63	0.63	0.63	0.61	0.64	0.64	0.64	0.63	0.64	0.63	0.64	0.62	0.63	0.64	0.63	0.61	0.66	0.66	0.66	0.65	0.62	0.61	0.62	0.60	0.52	0.51	0.52	0.48
RF	0.60	0.59	0.60	0.59	0.64	0.63	0.64	0.63	0.63	0.62	0.63	0.62	0.63	0.63	0.63	0.61	0.62	0.62	0.62	0.62	0.58	0.57	0.58	0.57	0.48	0.46	0.48	0.47
LR	0.44	0.40	0.44	0.41	0.55	0.53	0.55	0.53	0.53	0.50	0.53	0.49	0.53	0.51	0.53	0.51	0.52	0.47	0.52	0.48	0.49	0.45	0.49	0.45	0.44	0.43	0.44	0.30
KNN	0.57	0.56	0.57	0.56	0.60	0.60	0.60	0.60	0.57	0.56	0.57	0.56	0.58	0.58	0.57	0.57	0.56	0.57	0.56	0.58	0.56	0.58	0.57	0.48	0.45	0.48	0.46	
DT	0.60	0.59	0.60	0.59	0.64	0.63	0.64	0.63	0.63	0.62	0.63	0.62	0.63	0.63	0.63	0.61	0.58	0.58	0.58	0.54	0.54	0.54	0.48	0.45	0.48	0.46		
NB	0.49	0.51	0.49	0.46	0.54	0.55	0.54	0.53	0.50	0.50	0.50	0.49	0.50	0.51	0.50	0.47	0.52	0.50	0.52	0.51	0.49	0.49	0.49	0.47	0.42	0.44	0.40	
ADABOOST	0.49	0.51	0.49	0.47	0.56	0.62	0.56	0.57	0.63	0.61	0.63	0.61	0.60	0.60	0.60	0.57	0.61	0.60	0.61	0.60	0.59	0.60	0.58	0.51	0.48	0.51	0.47	
ET	0.60	0.59	0.60	0.59	0.64	0.63	0.64	0.63	0.63	0.62	0.63	0.62	0.63	0.63	0.63	0.61	0.61	0.60	0.61	0.60	0.57	0.56	0.57	0.48	0.46	0.48	0.46	
MLP	0.60	0.58	0.60	0.57	0.63	0.63	0.63	0.62	0.63	0.62	0.63	0.62	0.61	0.60	0.61	0.59	0.62	0.61	0.62	0.61	0.60	0.59	0.60	0.58	0.51	0.49	0.51	0.46
QDA	0.45	0.52	0.45	0.43	0.29	0.53	0.29	0.28	0.52	0.52	0.52	0.51	0.45	0.45	0.45	0.44	0.52	0.51	0.52	0.51	0.48	0.48	0.48	0.46	0.14	0.43	0.14	0.17
BAGGING	0.60	0.59	0.60	0.59	0.64	0.63	0.64	0.63	0.63	0.62	0.63	0.62	0.63	0.63	0.63	0.61	0.61	0.60	0.61	0.60	0.57	0.56	0.57	0.56	0.47	0.45	0.47	0.46

7 FEATURES	CHI-SQUARE				ANOVA				EXTRA_TREES				MUTUAL_INFORMATION				PERMUTATION IMPORTANCE				RANDOM_FOREST				RECURSIVE_FEATURE ELIMINATION			
MODEL	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SVM	0.55	0.51	0.55	0.52	0.57	0.54	0.57	0.55	0.52	0.46	0.52	0.47	0.53	0.52	0.53	0.50	0.56	0.53	0.56	0.53	0.50	0.45	0.50	0.46	0.45	0.36	0.45	0.32
GB	0.64	0.63	0.64	0.63	0.64	0.64	0.64	0.63	0.64	0.63	0.64	0.62	0.63	0.64	0.63	0.61	0.66	0.66	0.66	0.65	0.62	0.61	0.62	0.60	0.55	0.53	0.55	0.53
RF	0.59	0.58	0.59	0.59	0.64	0.63	0.64	0.63	0.60	0.59	0.60	0.60	0.63	0.64	0.63	0.61	0.64	0.64	0.64	0.64	0.59	0.58	0.59	0.58	0.51	0.50	0.51	0.51
LR	0.53	0.49	0.53	0.50	0.55	0.53	0.55	0.53	0.52	0.49	0.52	0.48	0.54	0.51	0.54	0.52	0.52	0.49	0.52	0.49	0.49	0.46	0.49	0.45	0.44	0.34	0.44	0.32
KNN	0.57	0.55	0.57	0.56	0.60	0.60	0.60	0.60	0.58	0.57	0.58	0.57	0.58	0.58	0.57	0.55	0.54	0.55	0.54	0.58	0.56	0.58	0.56	0.51	0.50	0.51	0.50	
DT	0.57	0.56	0.57	0.56	0.64	0.63	0.64	0.63	0.60	0.59	0.60	0.59	0.63	0.64	0.63	0.61	0.57	0.57	0.57	0.57	0.54	0.54	0.54	0.50	0.49	0.50	0.49	0.49
NB	0.50	0.51	0.50	0.47	0.54	0.55	0.54	0.53	0.51	0.50	0.51	0.49	0.50	0.51	0.50	0.47	0.52	0.50	0.52	0.51	0.48	0.48	0.48	0.46	0.44	0.44	0.44	0.42
ADABOOST	0.60	0.60	0.60	0.58	0.56	0.62	0.56	0.57	0.63	0.61	0.63	0.61	0.60	0.61	0.60	0.59	0.60	0.60	0.60	0.59	0.60	0.59	0.60	0.58	0.53	0.51	0.53	0.51
ET	0.58	0.57	0.58	0.57	0.64	0.63	0.64	0.63	0.60	0.59	0.60	0.59	0.63	0.64	0.63	0.61	0.63	0.62	0.63	0.62	0.57	0.56	0.57	0.57	0.51	0.50	0.51	0.50
MLP	0.61	0.60	0.61	0.59	0.62	0.63	0.62	0.60	0.62	0.61	0.62	0.61	0.62	0.63	0.62	0.61	0.62	0.61	0.62	0.61	0.60	0.58	0.60	0.58	0.54	0.53	0.54	0.52
QDA	0.43	0.52	0.43	0.43	0.29	0.53	0.29	0.28	0.52	0.52	0.52	0.51	0.46	0.53	0.46	0.44	0.52	0.51	0.52	0.51	0.48	0.48	0.48	0.46	0.32	0.45	0.32	0.28
BAGGING	0.58	0.58	0.58	0.58	0.64	0.63	0.64	0.63	0.60	0.59	0.60	0.60	0.63	0.64	0.63	0.61	0.62	0.61	0.62	0.62	0.57	0.56	0.57	0.56	0.50	0.49	0.50	0.50

10 FEATURES	CHI-SQUARE				ANOVA				EXTRA_TREES				MUTUAL_INFORMATION				PERMUTATION IMPORTANCE				RANDOM_FOREST				RECURSIVE_FEATURE ELIMINATION			
MODEL	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SVM	0.58	0.54	0.58	0.56	0.59	0.56	0.59	0.57	0.52	0.47	0.52	0.47	0.57	0.55	0.57	0.55	0.56	0.53	0.56	0.54	0.51	0.48	0.51	0.48	0.49	0.49	0.49	0.44
GB	0.68	0.67	0.68	0.67	0.67	0.66	0.67	0.66	0.64	0.63	0.64	0.62	0.64	0.64	0.64	0.63	0.67	0.66	0.67	0.66	0.64	0.63	0.64	0.56	0.55	0.56	0.55	
RF	0.64	0.63	0.64	0.63	0.67	0.66	0.67	0.66	0.61	0.60	0.61	0.61	0.64	0.63	0.64	0.63	0.65	0.64	0.65	0.64	0.61	0.60	0.61	0.51	0.51	0.51	0.51	
LR	0.55	0.50	0.55	0.52	0.56	0.54	0.56	0.54	0.52	0.49	0.52	0.49	0.53	0.51	0.53	0.51	0.54	0.51	0.54	0.51	0.52	0.50	0.52	0.50	0.47	0.45	0.47	0.40
KNN	0.58	0.57	0.58	0.57	0.62	0.62	0.62	0.62	0.59	0.58	0.59	0.58	0.61	0.60	0.61	0.60	0.56	0.55	0.56	0.55	0.59	0.57	0.59	0.57	0.51	0.51	0.51	0.50
DT	0.59	0.59	0.59	0.59	0.66	0.65	0.66	0.65	0.55	0.55	0.55	0.55	0.64	0.63	0.64	0.63	0.58	0.58	0.58	0.58	0.55	0.55	0.55	0.55	0.49	0.49	0.49	0.49
NB	0.53	0.54	0.53	0.52	0.55	0.56	0.55	0.54	0.52	0.52	0.52	0.50	0.54	0.54	0.54	0.52	0.48	0.49	0.48	0.46	0.52	0.52	0.52	0.50	0.46	0.47	0.46	0.44
ADABOOST	0.55	0.56	0.55	0.53	0.59	0.61	0.59	0.59	0.61	0.59	0.61	0.59	0.61	0.61	0.61	0.59	0.62	0.62	0.62	0.61	0.61	0.59	0.61	0.58	0.53	0.53	0.51	
ET	0.62	0.61	0.62	0.62	0.67	0.66	0.67	0.66	0.59	0.58	0.59	0.58	0.64	0.63	0.64	0.63	0.63	0.63	0.63	0.63	0.59	0.59	0.59	0.59	0.50	0.50	0.50	0.50
MLP	0.65	0.64	0.65	0.64	0.65	0.65	0.65	0.64	0.63	0.62	0.63	0.61	0.64	0.63	0.64	0.63	0.64	0.62	0.64	0.63	0.63	0.62	0.63	0.62	0.55	0.54	0.55	0.54
QDA	0.49	0.57	0.49	0.46	0.49	0.58	0.49	0.50	0.53	0.53	0.53	0.51	0.12	0.53	0.12	0.15	0.45	0.48	0.45	0.44	0.52	0.53	0.52	0.50	0.41	0.46	0.41	0.36
BAGGING	0.62	0.61	0.62	0.62	0.67	0.66	0.67	0.66	0.59	0.58	0.59	0.58	0.64	0.63	0.64	0.63	0.63	0.63	0.63	0.63	0.59	0.59	0.59	0.59	0.50	0.50	0.50	0.50

Table 7.3: Top 5,6,7 & 10 features

The comparative analysis conducted for the model performance of 5, 6 and 7 features against over 11 features reveals noteworthy insights into the effect of feature selection on the prediction performance and capabilities of diverse machine learning algorithms.

As displayed in Table 7.3, when employing over 11 features, GB consistently produced the highest accuracy across all the feature selection methods, ranging from 0.68 to 0.69 in both test and validation datasets. However, SVM engendered varying performance values based on the feature selection method utilized, with values ranging from 0.46 to 0.61. RF and ET also produced consistent accuracy, ranging from 0.65 to 0.69. Conversely, the Naive Bayes (NB) model exhibited a significant increase in accuracy without the application of feature selection versus using over 11 features from the various features selection (FS) methods, ranging from a low of 0.20 (without FS) to 0.54 (with FS).

In comparison to the performance of the models using 5 features, it was noted that the accuracy metrics of the models generally revealed a decline when the number of features was reduced. For instance, the SVM which had an accuracy of 0.59 with over 11 features, reduced to 0.56 when limited to 5 features. This trend is recurrent across multiple models, including GB, RF, LR, and KNN among others. The reduced feature set appeared to have a specifically adverse effect on the performance of complex models such as MLP and GB, where the accuracy declined from 0.68 to 0.61 and 0.69 to 0.63, respectively. Precision, recall, and F1 score metrics also show a similar decline in performance across models. The decrease in the number of features presumably results in a loss of correlating features, affecting the model's ability to effectively capture the intricacies of the underlying patterns within the dataset. Thus, the trade-off between model complexity and the number of features must be prudently considered, as incessant feature reduction may compromise the performance of the model. Figure 7.9 shows the model performance using the top 5 features of varying feature selection methods.

5 FEATURES

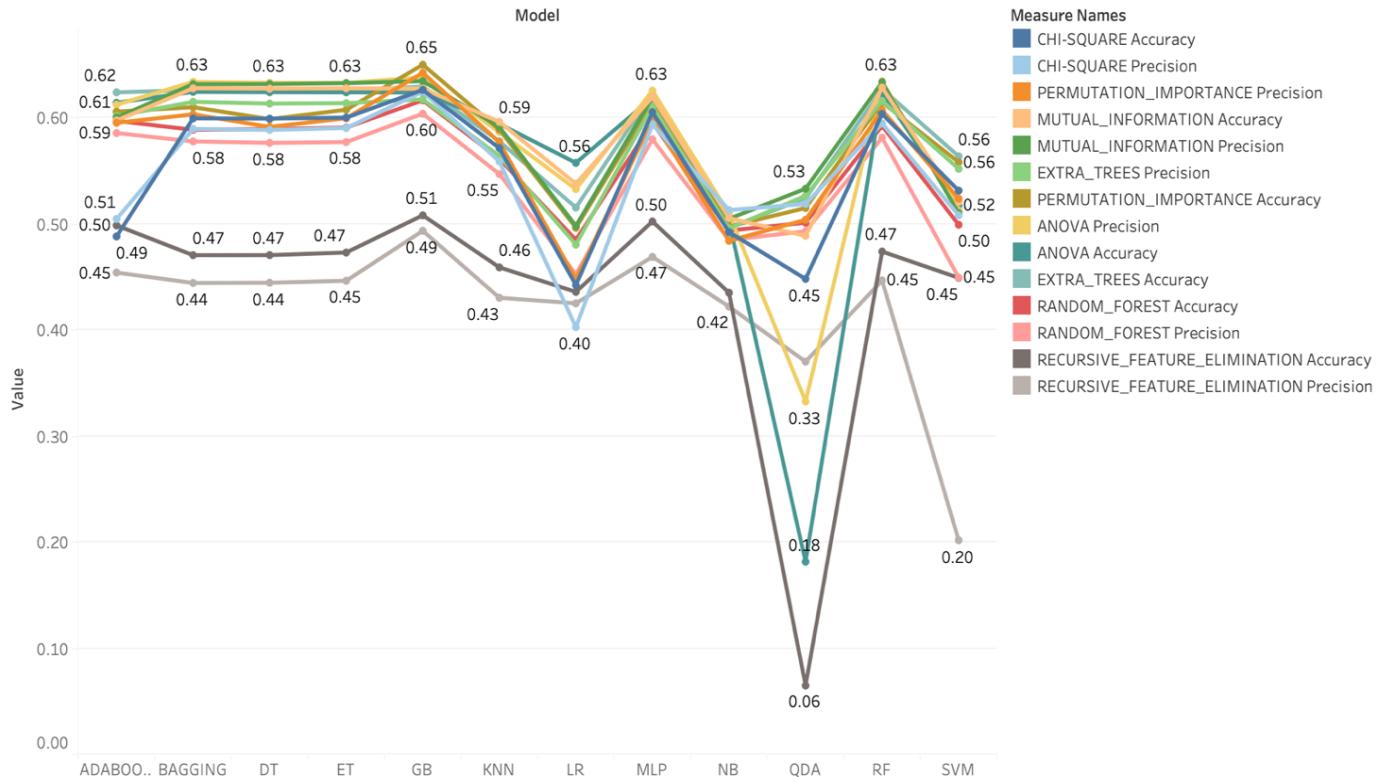


Figure 7.9: Model performance using top 5 features of varying feature selection method

Similarly, Figure 7.10 shows the model performance using the top 6 features of varying feature selection methods. The model's results with the 6 most relevant features demonstrate a similar form of stability in model performance. GB maintains its position as the top-performing model with consistent accuracies around 0.63. However, using ANOVA, the addition of one important feature showed an increase in performance from 0.62(5 features) to 0.64(6 features). Nonetheless, RF and ET also demonstrated stable performance, with accuracy values ranging from 0.60 to 0.63. SVM also improved compared to using just 5 features, ranging from 0.52 to 0.57.

6 FEATURES

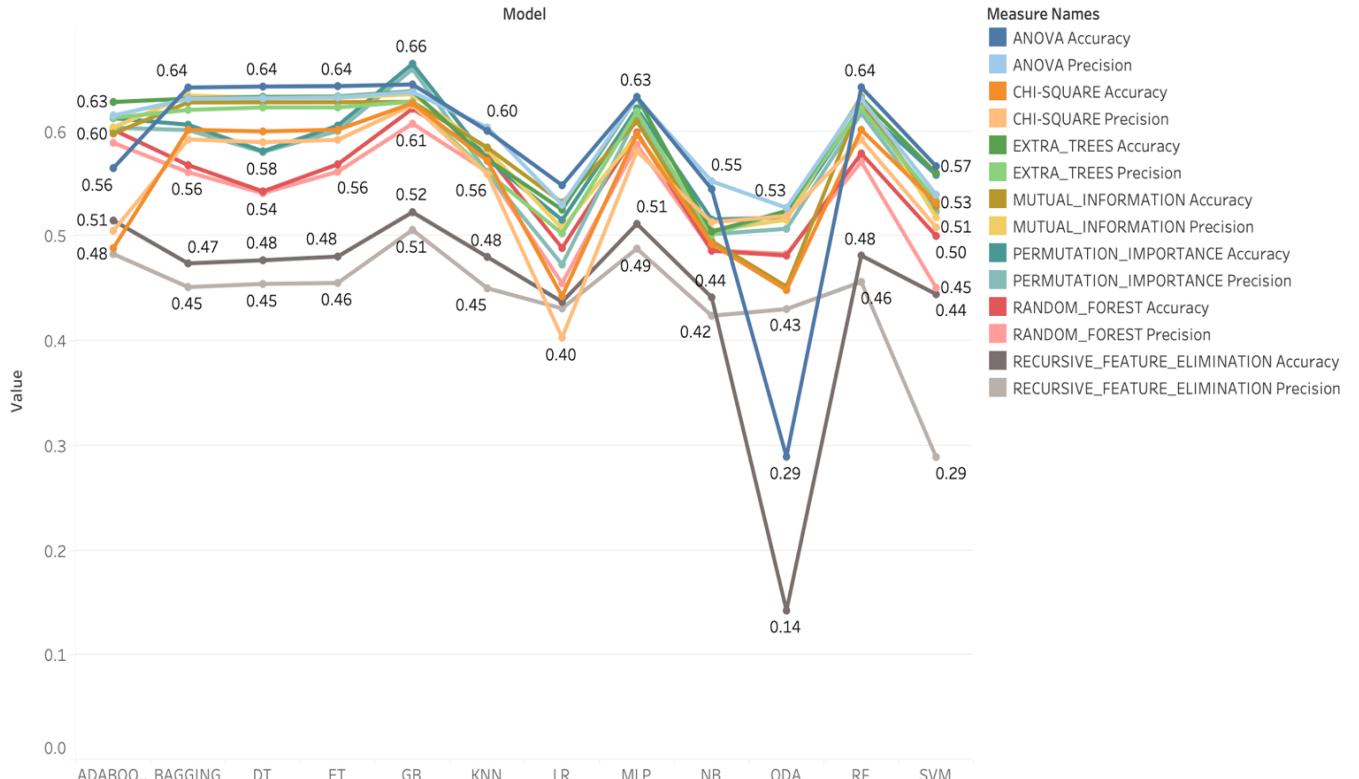


Figure 7.10: Model performance using top 6 features of varying feature selection method.

SVM establishes enhanced stability compared to the 6-feature scenario. However, SVM experienced a decrease in performance using the ET feature selection method, from 0.56 to 0.52. This suggests that the model still shows sensitivity to different feature sets, denoting the necessity for meticulous consideration in feature selection for optimal SVM performance. Also, NB displays a slight improvement in performance compared to the 6-feature scenario, achieving up to 0.54. Adaboost and MLP models display consistent performance, with performance accuracy ranging from 0.58 to 0.61. These algorithms appear to benefit from the added information delivered by the 7 important features, maintaining stable accuracy levels. GB projects as a resilient algorithm, maintaining good accuracy across different feature sets. RF and ET also prove effective in handling a moderately higher number of features.

It is evident that majority of the model benefits from the addition of features and engenders better performance. Therefore, the choice of a number of features significantly influences the model performance, accentuating the importance of cautious feature selection to strike a balance between model complexity and predictive accuracy. In which case, based on this increase in performance by the addition of features, a 10-feature scenario was further explored. Figure 7.11 shows the model performance using the top 7 features of varying feature selection methods.

7 FEATURES

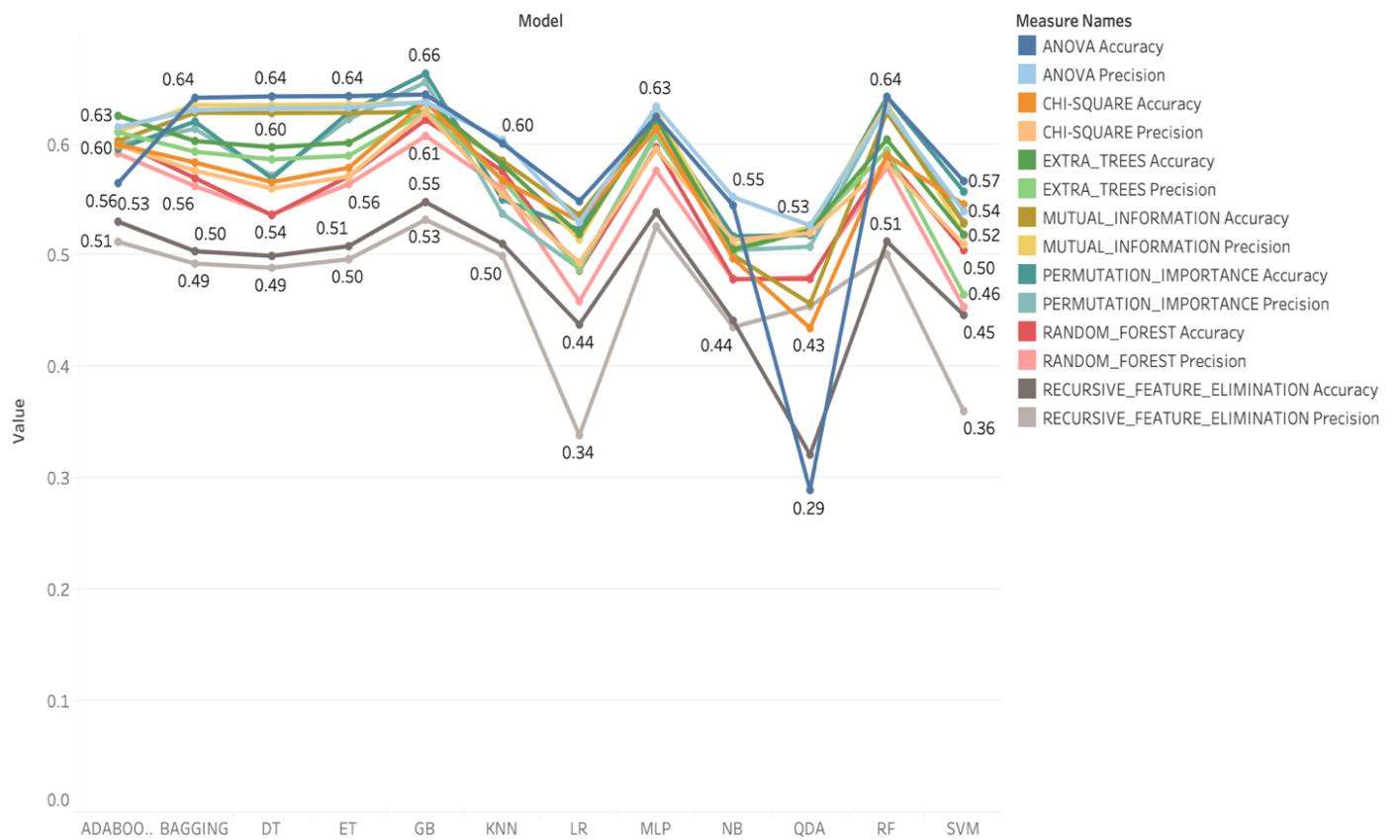


Figure 7.11: Model performance using top 7 features of varying feature selection method

Figure 7.12 shows the model performance using the top 10 features. The utilization of the 10 most important features provides more insights into how machine learning models perform with increased feature dimensionality. GB maintains good consistency and remains at the top position with performance accuracy values ranging from 0.63 to 0.68. Using 10 features, the highest performance value for GB is 0.68 which is a slim margin to the accuracy values achieved using 17 features produced by chi-square. RF and Extra ET also perform good outcomes with a high of 0.67, denoting their robustness in handling a relatively larger set of features. SVM portrays improved stability in comparison to scenarios with a lesser number of features, producing an accuracy between 0.47 and 0.59. This potentially suggests that SVM benefits from a more comprehensive set of features. However, it still faces challenges in producing higher accuracy.

KNN reaches relatively stable performance, with accuracy values from 0.57 to 0.62. The model also appears to benefit from the increased features. Decision Tree (DT) performance remains variable and shows that it benefits from additional features, however, it still faces consistency issues across different feature sets. Also, a significant increase in the feature set does not elicit

better performance. For example, without the application of feature selection DT achieved 0.58 and 0.66 using 10 important features produced by ANOVA. On the other hand, NB exhibits stability in accuracy and benefits from an increase in the set of features. However, its performance appears limited in comparison to more complex algorithms. MLP shows its capacity to handle a larger set of features with the production of competitive accuracy values. QDA experiences issues in handling a larger number of features. For example, using 5 sets of features from the MI feature selection method achieved 0.49 while using 10 feature sets and without feature selection produced 0.12 and 0.13 respectively. The significant drop in accuracy implies that QDA may not be well-matched for scenarios with a more extensive set of features. Ensemble methods, such as Bagging, showcase robust performance and competitive results. These methods show their efficacy in leveraging the collective strength of various models. Considering the good performance of chi-square, the 10 top features were selected for subsequent development.

10 FEATURES

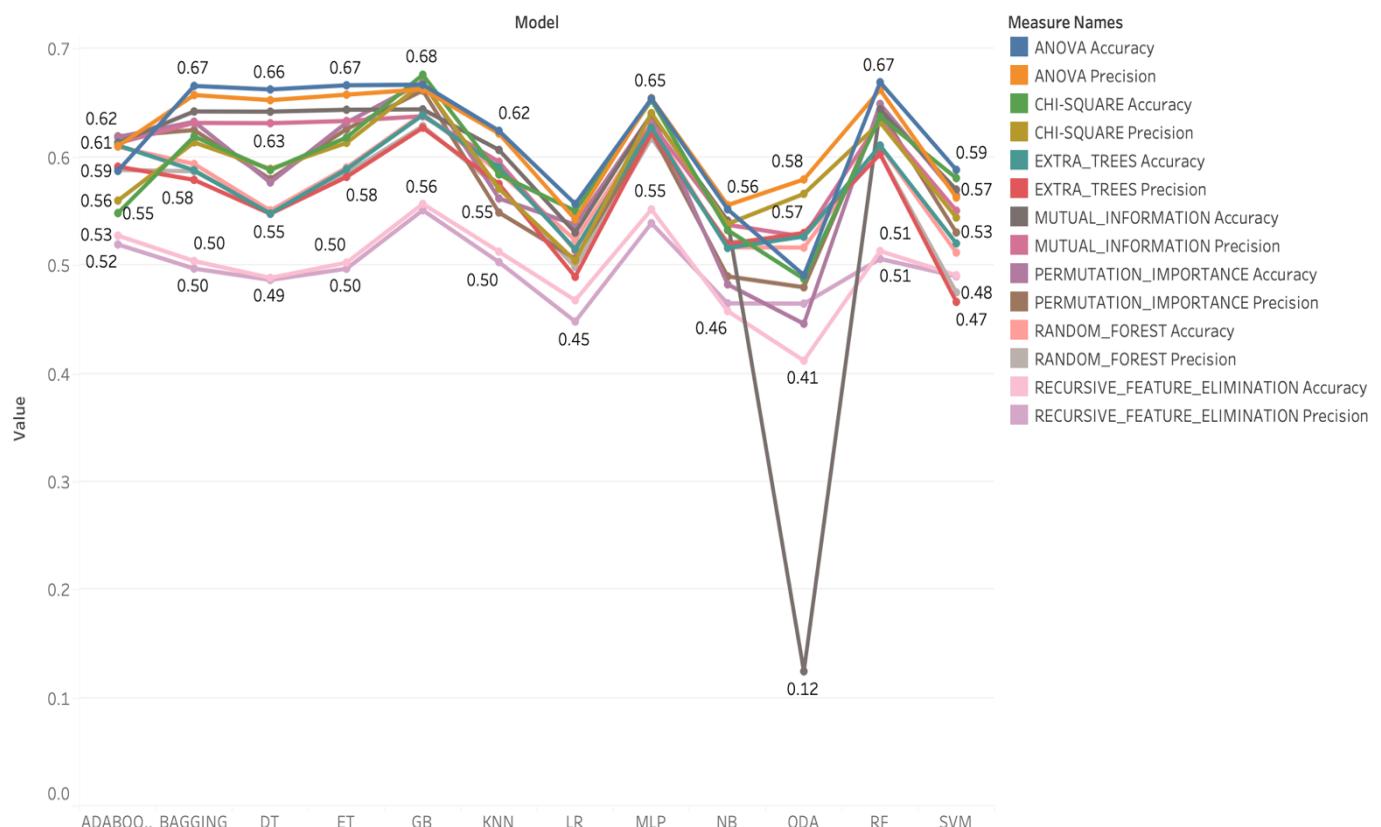


Figure 7.12: Model performance using top 10 features varying feature selection method

7.8.7.2 Impact on Regression Model Performance

Similarly, for regression, the same experiment of systematic assessment of algorithms predictive capabilities using 5, 6, and 7 of the most important features deduced by the different

feature selection method. Table 6.8 below shows the model performance using 5, 6 and 7 most relevant features of diverse feature selection method.

It shows the different performance metric namely R-squared, mean absolute error (MAE), and root mean squared error (RMSE). The result shows that the models generally achieved higher R-squared values and lower errors when consuming over 11 features compared to only 5 features. For example, Random Forest (RF) model achieved a R-squared value of 0.86 with over 11 MI features, it decreases in performance to 0.51 using 5 features. Similarly, Gradient boosting (GT) displays a higher R-squared of 0.85 with over 11 MI features compared to 0.56 using 5 features. While the utilization of over 11 features may lead to improved performance, it engenders potential risk of overfitting. The reduced set of features may essentially help mitigate overfitting by focusing on the most relevant feature. This is apparent in cases such as ADABOOST, where the R-squared increases from 0.64 with over 11 features to 0.68 with 5 features.

Subsequently, the result using 6 features showed minimal difference performance over results using 5 relevant features. However, not as good at results using over 11 features. For example, DNN produced 0.49 and 0.50 R-squared values using 5 and 6 features respectively. For 7 features scenario, GB and RF maintain strong performance with 7 features, consistently achieving high R-squared values and relatively low errors. SVR exhibits improved performance with 7 features compared to the scenarios with 5 and 6 features.

Table 7.4: Top 5,6,7 & 10 features

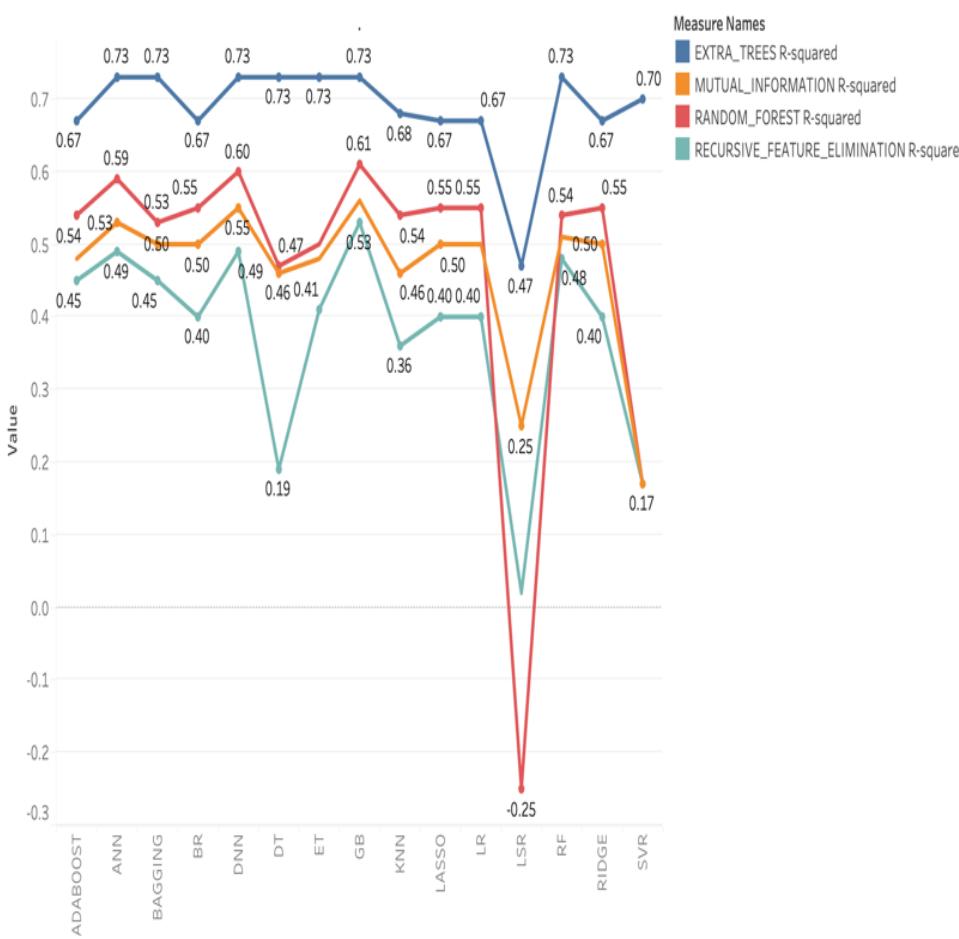
5 FEATURES		EXTRA TREES			MUTUAL INFORMATION			RANDOM FOREST			RECURSIVE FEATURE ELIMINATION		
Model	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE	
SVR	0.70	67.73	93.69	0.17	112.71	154.39	0.17	113.09	154.72	0.17	113.17	155.08	
ANN	0.73	65.29	88.43	0.53	84.11	115.95	0.59	80.02	108.56	0.49	90.27	121.66	
GB	0.73	64.68	87.59	0.56	81.96	112.44	0.61	77.60	105.91	0.53	85.40	116.22	
RF	0.73	64.91	88.03	0.51	86.59	118.70	0.54	82.97	115.37	0.48	88.08	122.02	
LR	0.67	74.38	98.13	0.50	87.16	119.75	0.55	85.51	113.41	0.40	99.85	131.99	
KNN	0.68	71.05	95.34	0.46	90.59	124.98	0.54	84.06	115.71	0.36	97.71	135.63	
DT	0.73	65.09	88.40	0.46	90.78	124.99	0.47	87.56	123.51	0.19	107.90	152.86	
ADABOOST	0.67	73.77	97.84	0.48	92.22	122.72	0.54	86.01	115.36	0.45	95.79	126.18	
BAGGING	0.73	64.96	88.07	0.50	87.35	120.02	0.53	83.25	116.01	0.45	90.26	125.41	
ET	0.73	65.09	88.40	0.48	88.98	122.71	0.50	85.79	120.17	0.41	93.41	130.39	
RIDGE	0.67	74.36	98.13	0.50	87.16	119.75	0.55	85.51	113.41	0.40	99.85	131.99	
LASSO	0.67	74.19	98.09	0.50	87.07	119.76	0.55	85.33	113.37	0.40	99.87	132.00	
BR	0.67	74.36	98.12	0.50	87.16	119.74	0.55	85.51	113.41	0.40	99.86	131.99	
LSR	0.47	80.87	123.57	0.25	101.67	146.97	-0.25	133.19	189.80	0.02	118.62	168.02	
DNN	0.73	65.36	88.94	0.55	83.10	113.68	0.60	78.98	107.78	0.49	88.77	121.46	
6 FEATURES		EXTRA TREES			MUTUAL INFORMATION			RANDOM FOREST			RECURSIVE FEATURE ELIMINATION		
Model	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE	
SVR	0.70	67.63	93.52	0.17	112.73	154.41	0.17	113.03	154.62	0.17	113.18	155.10	
ANN	0.73	65.53	88.69	0.52	83.33	117.99	0.59	80.11	108.51	0.43	97.16	128.07	
GB	0.73	64.68	87.59	0.56	81.80	112.14	0.61	77.60	105.91	0.54	84.71	115.00	
RF	0.73	64.94	88.09	0.52	86.20	117.89	0.54	83.21	115.55	0.53	84.43	116.91	
LR	0.67	74.39	98.14	0.50	86.77	119.62	0.55	85.51	113.41	0.40	99.82	131.96	
KNN	0.68	71.12	95.33	0.46	90.83	125.24	0.54	83.94	115.75	0.37	96.19	134.38	
DT	0.73	65.09	88.40	0.45	91.15	125.60	0.47	87.46	123.21	0.23	105.55	149.08	
ADABOOST	0.67	74.29	97.81	0.47	93.87	123.44	0.53	87.35	116.34	0.45	96.63	126.42	
BAGGING	0.73	64.95	88.06	0.51	86.95	119.13	0.53	83.26	116.30	0.49	87.29	120.93	
ET	0.73	65.10	88.41	0.47	88.84	123.06	0.50	85.80	120.05	0.46	89.42	125.34	
RIDGE	0.67	74.36	98.13	0.50	86.77	119.62	0.55	85.51	113.41	0.40	99.82	131.96	
LASSO	0.67	74.19	98.09	0.51	86.66	119.43	0.55	85.33	113.37	0.40	99.86	131.99	
BR	0.67	74.36	98.12	0.50	86.76	119.61	0.55	85.51	113.41	0.40	99.82	131.96	
LSR	0.49	80.48	121.18	0.31	98.08	141.55	-0.38	138.96	199.34	0.02	118.93	167.71	
DNN	0.73	65.43	88.42	0.49	83.02	121.46	0.60	78.79	107.74	0.50	87.94	119.93	
7 FEATURES		EXTRA TREES			MUTUAL INFORMATION			RANDOM FOREST			RECURSIVE FEATURE ELIMINATION		
Model	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE	
SVR	0.71	64.44	91.36	0.17	112.73	154.42	0.18	112.52	154.04	0.24	108.19	148.54	
ANN	0.76	60.70	83.63	0.53	82.41	116.75	0.73	64.66	88.38	0.53	86.75	116.19	
GB	0.77	59.18	81.55	0.57	80.77	111.11	0.75	61.55	84.26	0.64	75.11	102.27	
RF	0.76	59.57	82.86	0.52	85.08	117.40	0.71	66.00	91.63	0.64	72.16	102.02	
LR	0.68	72.97	95.96	0.50	86.78	119.72	0.68	72.59	96.45	0.48	92.40	122.57	
KNN	0.72	64.88	89.95	0.48	88.24	122.15	0.69	68.35	94.98	0.45	89.37	125.98	
DT	0.76	59.92	83.50	0.44	91.33	126.69	0.64	72.33	101.35	0.35	94.42	136.77	
ADABOOST	0.66	77.79	99.53	0.47	93.23	123.61	0.67	76.05	97.40	0.45	102.86	125.90	
BAGGING	0.76	59.57	83.00	0.50	86.41	119.57	0.70	66.99	93.13	0.61	74.65	105.46	
ET	0.76	59.84	83.33	0.48	88.06	122.73	0.67	69.53	96.98	0.61	74.21	105.73	
RIDGE	0.68	72.95	95.95	0.50	86.78	119.72	0.68	72.59	96.45	0.48	92.40	122.57	
LASSO	0.68	72.80	95.92	0.51	86.66	119.47	0.68	72.41	96.42	0.48	92.41	122.60	
BR	0.68	72.94	95.95	0.50	86.77	119.72	0.68	72.59	96.45	0.48	92.41	122.57	

LSR	0.54	75.63	114.94	0.28	99.66	143.71	-0.01	117.38	170.72	0.11	113.77	160.01
DNN	0.77	59.74	82.31	0.50	84.48	120.64	0.74	63.30	86.79	0.54	84.97	115.13
10 FEATURES												
Model	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE	R-squared	MAE	RMSE
SVR	0.18	112.10	153.46	0.18	112.23	153.85	0.18	112.48	154.02	0.24	108.10	148.28
ANN	0.81	52.88	74.24	0.73	61.20	87.84	0.77	57.04	80.71	0.56	83.18	112.65
GB	0.83	49.39	69.99	0.77	56.37	81.95	0.83	51.26	72.60	0.64	74.40	101.46
RF	0.81	52.20	74.15	0.72	61.02	90.27	0.81	50.86	73.22	0.66	69.90	99.13
LR	0.71	69.48	92.11	0.66	72.75	98.35	0.70	70.31	93.06	0.48	92.49	122.46
KNN	0.75	59.69	85.47	0.70	64.05	93.77	0.74	60.33	86.49	0.43	91.07	127.86
DT	0.69	66.58	94.11	0.59	72.39	108.17	0.67	67.11	98.13	0.36	94.18	136.17
ADABOOST	0.62	86.87	105.17	0.64	78.55	102.54	0.62	86.91	104.66	0.38	111.92	134.26
BAGGING	0.80	53.75	76.31	0.71	62.23	91.99	0.80	53.15	76.29	0.63	72.45	102.99
ET	0.78	55.88	79.51	0.67	64.72	96.87	0.80	52.37	75.81	0.63	72.29	103.07
RIDGE	0.71	69.43	92.09	0.66	72.75	98.35	0.70	70.31	93.06	0.48	92.49	122.46
LASSO	0.71	69.27	92.10	0.67	72.61	98.25	0.70	70.19	93.09	0.48	92.52	122.53
BR	0.71	69.43	92.09	0.66	72.74	98.34	0.70	70.31	93.06	0.48	92.50	122.47
LSR	-0.13	119.29	180.31	0.16	105.40	155.71	0.25	98.74	146.96	0.10	115.41	161.53
DNN	0.82	50.88	71.39	0.71	63.67	91.93	0.80	53.79	76.10	0.56	81.24	112.72

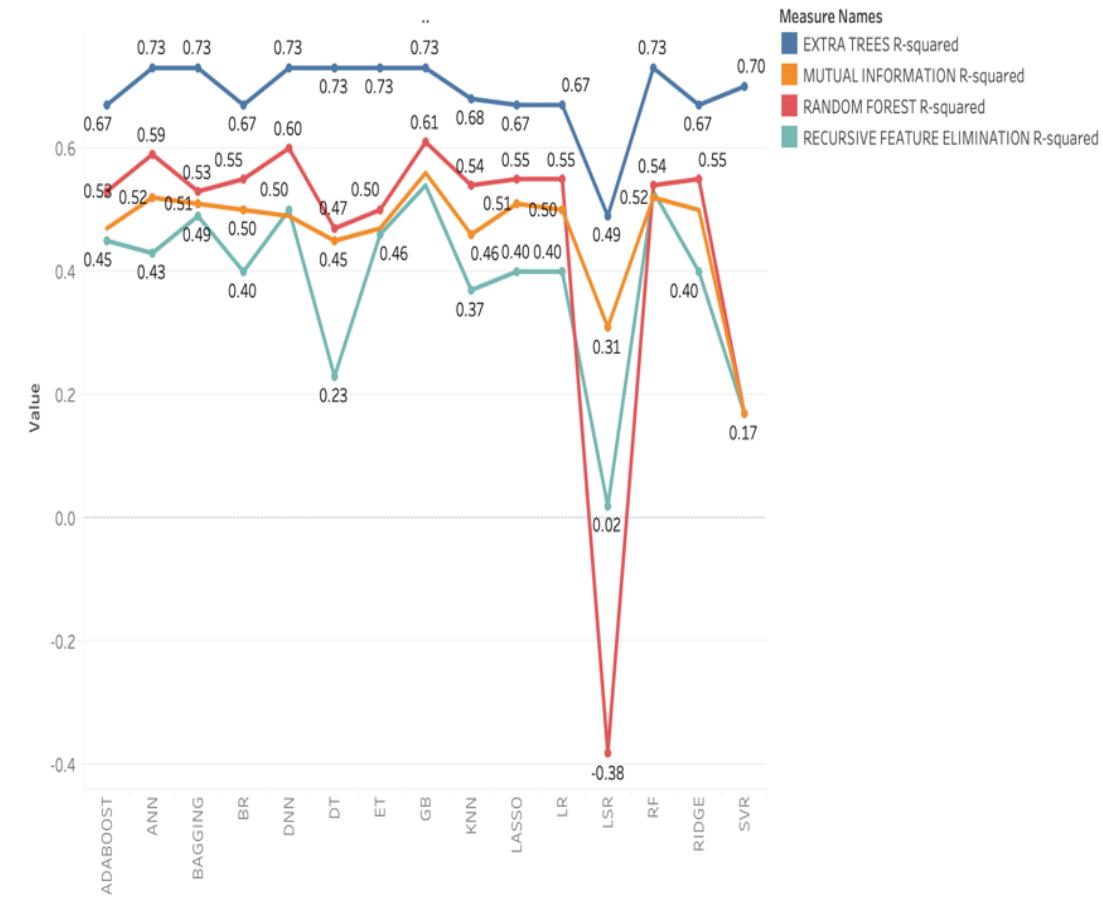
Furthermore, using 10 important features, several models demonstrated improved performance. Gradient Boosting (GB) continue to demonstrate strong prediction performance, achieving a notable R-squared of 0.83 and only 10 features were initially selected as most relevant. Considering the good performance of random forest feature selection, the 10 top features were selected for subsequent development.

Figure 7.13 shows the trend in model performance values using 5,6,7 and 10 features.

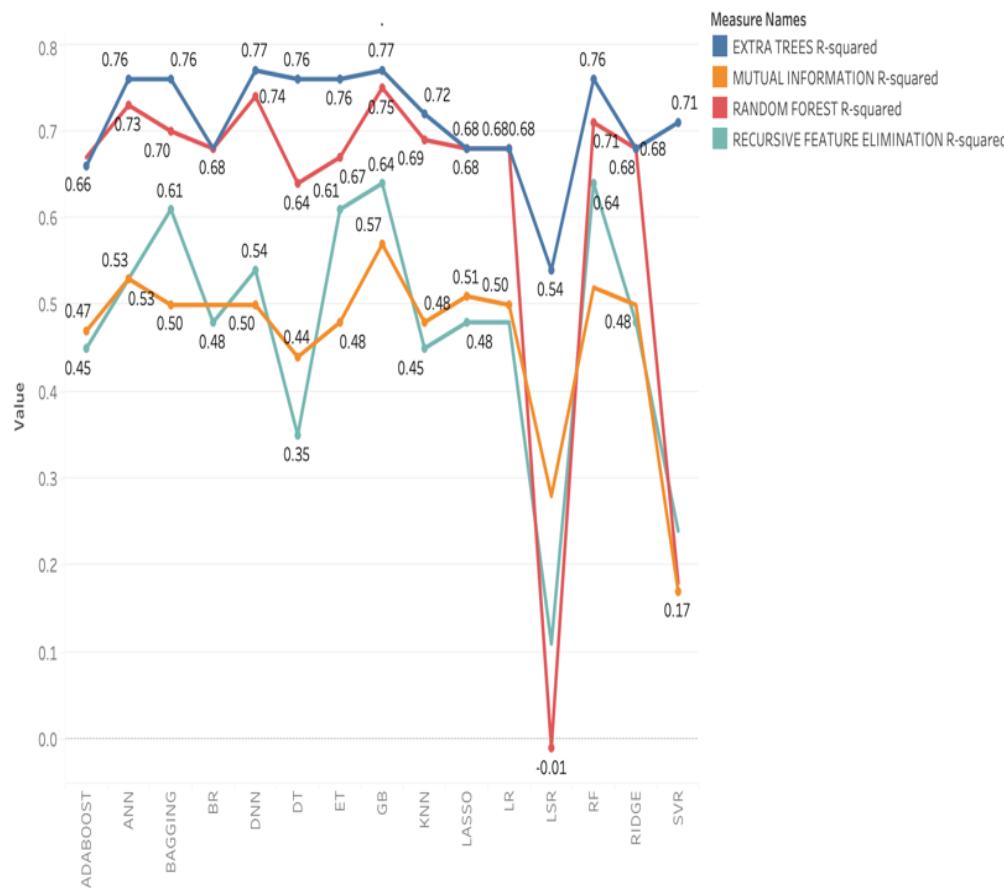
5 FEATURES



6 FEATURES



7 FEATURES



10 FEATURES

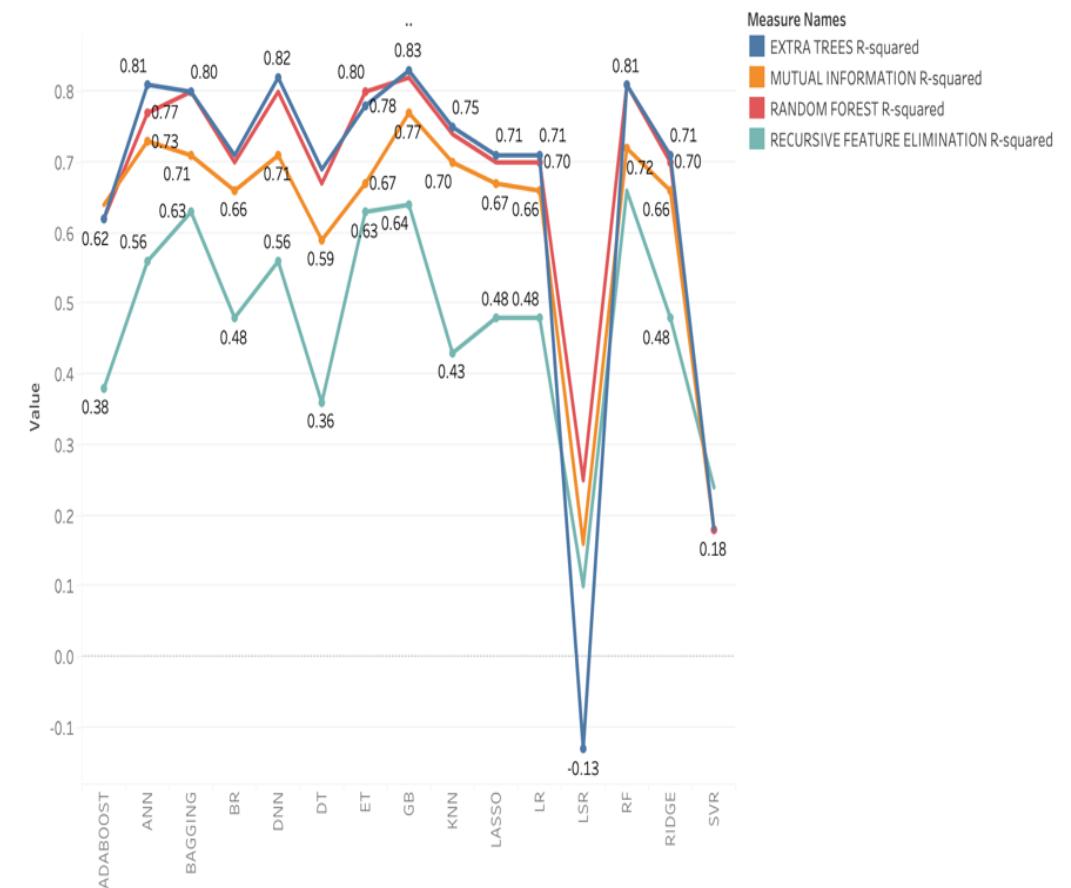


Figure 7.13: Model performance using top 5,6,7 & 10 features of varying feature selection method.

7.8.8 H6: Machine Learning Algorithms produce better performing Building Energy Prediction Performance Models than Statistical Method.

Many studies have conducted a comparative analysis of the performance of different Statistical, or AI/ML-based tools for energy consumption prediction(Ahmed Gassar et al., 2019a; Alduailij et al., 2021; S. Cho et al., 2019; Liao et al., 2020; Lin et al., 2021; Parhizkar et al., 2021). However, AI and Statistical based tools have at different times outperformed each other. For example, Somu et al., (2020) compared SVM(ML) and ARIMA(Statistical) for predicting energy consumption and it was noted that SVM produced better performance.

As shown in the framework in Figure 7.3, the phases 1 to 5 remain the same and all models were developed for this comparison. The box plot below (Figure 7.14) helps to reveal insights into consistency and variability of their predictions. Models with narrower boxes and shorter whiskers denote more stable and consistent performance across the datasets, while wider boxes and longer whiskers imply higher variability. Therefore, the box plot displays models that consistently perform well (i.e., ET, RF and GB) and those comprising of more variability in their predictions (Adaboost, LSR). This comparative analysis is essential for a more informed selection of the most reliable model based on both performance and consistency for your specific application.

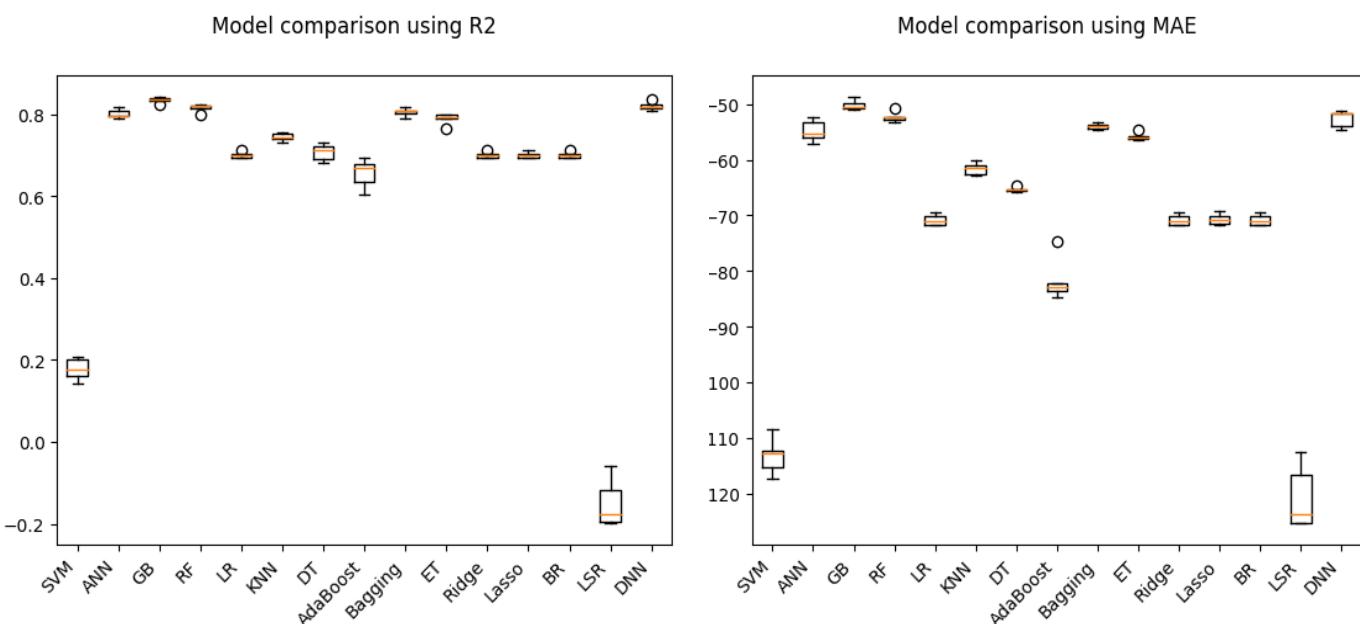


Figure 7.14: Model comparison using R^2 and MAE

More elaborately, Figure 7.14 shows a clear comparison as the R2 box plot displays GB interquartile range is the closest to 1.0. The plot shows no significant difference between GB and RF-based models in the R2 box plot. However, Table 6.8 clarifies this ambiguity of insignificant variation between GB and DNN with an R2 of 0.83 and 0.82 respectively. Furthermore, the worst predictive model is LSR as it is closest to 0.0 for R2 judging by the interquartile range and it is the farthest from 0.0 for the MAE box plot. LSR and SVM are considered the most inaccurate predictive models based on the significant variation to the good predictive models ET, DNN, RF and GB across all performance measures. Similar to the comparative study by Guo et al., (2018), LR(Statistical) also outperforms SVM for predicting energy consumption in this research. Machine learning models have consistently achieved better performance than statistical models and a careful review of these studies shows that ML produce better performance in majority of the studies (see chapter 4). Therefore, the following hypothesis is formulated.

- **Null hypothesis H_0 :** Machine learning algorithm produces a better building energy prediction performance model than the statistical method.
- **Alternative hypothesis H_A :** Machine learning algorithms do not produce better building energy prediction performance models than statistical methods.

Given the good performance of the machine learning models, particularly Gradient Boosting (GB), Random Forest (RF), and Deep Neural Network (DNN), in both classification and regression tasks, it is equitable to accept the null hypothesis. The null hypothesis posits that machine learning algorithms produce better building energy prediction performance models than statistical methods. The evidence from the observed high r-squared value of GB, RF, and ET across regression tasks which affirms the perception that machine learning algorithms are indeed effective in building energy prediction models.

7.8.9 H7: Deep Learning Algorithm outperforms Classical Machine Learning Algorithms.

In recent years, deep learning algorithms have shown great promise in various fields (Abrol et al., 2021; Brinker et al., 2019; Hekler et al., 2019). It has become prominent based on the

production of good performance in various other studies (Almalaq and Zhang, 2019; C. Fan et al., 2017; Somu et al., 2021). The null and alternative hypothesis are stated as follows:

- **Null hypothesis H_0 :** Deep learning algorithm outperforms classical machine learning algorithms.
- **Alternative hypothesis H_A :** Deep learning algorithm do not outperform classical machine learning algorithms.

The comparison of regression models in Table 7.2 without feature selection reveals that the Deep Neural Network (DNN) outperforms other models, with the exception of Extra trees, Bagging and GB. Similarly, [Sadeghi et al., \(2020\)](#) a comparative analysis of different machine learning algorithms and Deep neural networks (DNN) produced the best performance. However, this performance is further to the optimization of the model. Contextually, given that only DNN was employed in this research considering the nature of the data, other studies have employed time series models like Long Short-Term Memory (LSTM) for similar tasks([da Silva et al., 2022](#); [Kong et al., 2019](#); [X. Shao et al., 2020](#)), it becomes important to consider the limitations of the current study. The alternative hypothesis may be accepted with caution, as the absence of other deep learning architectures, such as LSTM, in the comparison limits the generalizability of the findings.

Deep neural networks, particularly DNNs, are powerful models known for their ability to capture complex relationships in data. The observed superior performance of DNN in comparison to classical machine learning algorithms is a promising result. Howbeit, while the results show promising performance with DNN, it would be prudent to recognize the study's limitations and acknowledge the need for additional exploration of other deep learning approaches, especially in the context of time series modelling for building energy prediction. Consequently, the alternative hypothesis is accepted.

7.8.10 Reliability Analysis (Data Size)

This research conducts a reliability analysis to investigate the effect of sample size on classification and regression performance. Five different percentages (i.e., 20%, 40%, 60%, 80% and 100%) of data were utilized as training and testing sets to develop energy use rating prediction models. This estimation approach will adopt ten machine learning algorithms, some

of which have been applied in energy use prediction namely Artificial Neural Network (ANN) (Ahmad et al., 2014, 2017a; Li and Yao, 2020), Gradient Boosting (GB)(Cheng Fan et al., 2017; Wang et al., 2020), K Nearest Neighbour (KNN)(Wang et al., 2020), Deep Neural Network (DNN) (Kadir Amasyali and El-Gohary, 2021; Lei et al., 2021), Random Forest (RF)(Ahmad et al., 2017a; Z. Wang et al., 2018b), Decision tree (DT) (Chou and Bui, 2014; Wang et al., 2020; Yu et al., 2010) Stacking, Support Vector Machine (SVM)(Dong et al., 2021a; Niu et al., 2010; Wang et al., 2020), and Linear Regression (LR)(Chou and Bui, 2014). This research explores the utilization of prediction methods for annual energy consumption using an increased dataset.

Furthermore, this chapter incorporates big data analytics to train and evaluate the models not only based on accuracy but also in terms of computational efficiency. Though accuracy is a key measure, the efficiency of these models is also paramount given the goal of the developer and especially when dealing with large datasets. The utilization of big data analytics aims to provide a holistic understanding of the performance of models, considering both accuracy and computational efficiency. Given that the model's effectiveness goes beyond its ability to make accurate predictions; it must also demonstrate efficiency in handling substantial amounts of data(Alaka et al., 2018; Bourhnane et al., 2020; Olu-Ajayi et al., 2022b). The schematic diagram of this chapter is displayed in Figure 7.15 below however the phases 1 to 4 in Figure 7.3 remain the same.

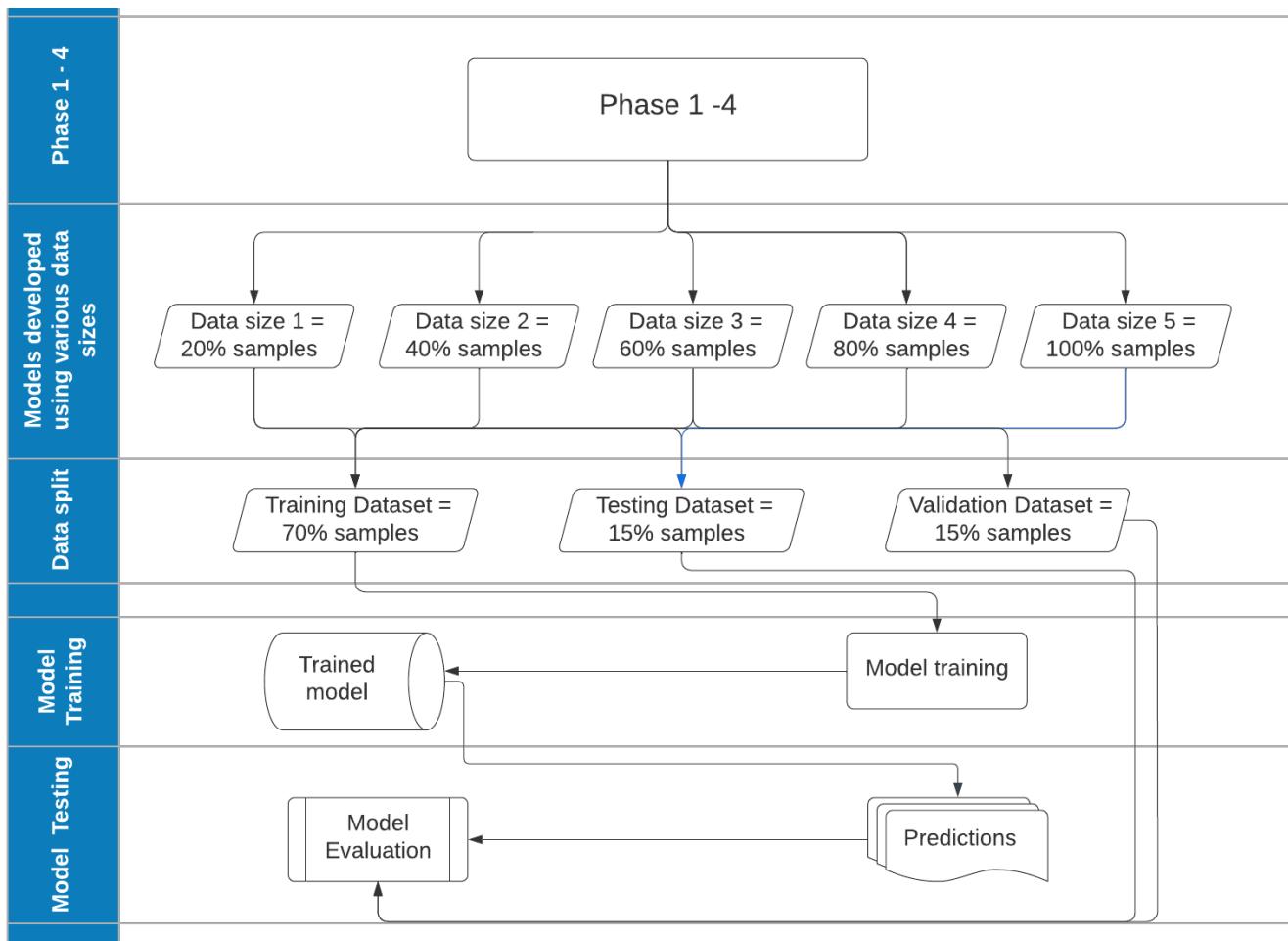


Figure 7.15: Flowchart diagram of the prediction framework

7.8.10.1 Data Size Impact on Classification Model Performance

The most effective predictive model across five cases of data availability (ranging from 20% to 100%) using the model performance measures provides valuable insights into their robustness and scalability. The performance values of accuracy closest to one are considered the most effective model. As summarized in Table 7.17, the Support Vector Machine (SVM) shows consistent accuracy around 0.59, precision, recall, and F1 scores between 0.58 and 0.59 across the various dataset sizes, denoting a stable performance. The gradient Boosting (GB) and Random Forest (RF) models demonstrate resilience and produce high accuracy (around 0.69), precision, recall, and F1 scores, along with impressive AUC scores.

The Linear Regression (LR) model shows gradual improvement in performance as the data size increases, signifying its sensitivity to the amount of training data. K Nearest Neighbour (KNN) and Decision Tree (DT) models showed reliable performance, with KNN exhibiting minor

improvements with larger datasets. Naive Bayes (NB) and Adaboost showed sensitivity to increases in data size, with fluctuations in performance metrics. Extra Trees (ET) and Multi-Layer Perceptron (MLP) consistently produced good performance around 0.66, showcasing their ability to handle varying amounts of data effectively. However, Quadratic Discriminant Analysis (QDA) and Bagging demonstrate mixed performance, suggesting potential sensitivity to data size.

The change in data size had effects on the performance of the different algorithms. It is noted that Gradient Boosting (GB) outperformed other models using 20% data and there appear no significant changes in the prediction performance (accuracy) with the increased sample size. This suggests that the size of the data has no direct impact on the predictive accuracy. Therefore, 20% and larger data can be considered sufficient for energy use prediction using Gradient Boosting (GB).

Table 7.5 below, with the most effective model in terms of performance measures indicated in bold.

CLASSIFICATION	20% DATA					40% DATA					60% DATA					80% DATA					100% DATA				
MODEL	Accuracy	Precision	Recall	F1	AUC Score	Accuracy	Precision	Recall	F1	AUC Score	Accuracy	Precision	Recall	F1	AUC Score	Accuracy	Precision	Recall	F1	AUC Score	Accuracy	Precision	Recall	F1	AUC Score
SVM	0.59	0.57	0.59	0.57		0.58	0.55	0.58	0.56		0.59	0.56	0.59	0.57		0.59	0.57	0.59	0.57		0.59	0.57	0.59	0.57	
GB	0.68	0.67	0.68	0.67	0.89	0.69	0.68	0.69	0.68	0.89	0.69	0.69	0.69	0.69	0.88	0.69	0.68	0.69	0.68	0.90	0.69	0.68	0.69	0.68	0.90
RF	0.68	0.68	0.68	0.67	0.88	0.68	0.67	0.68	0.67	0.89	0.69	0.69	0.69	0.68	0.88	0.68	0.68	0.68	0.68	0.90	0.69	0.69	0.69	0.68	0.90
LR	0.55	0.51	0.55	0.52	0.71	0.55	0.51	0.55	0.51	0.72	0.56	0.52	0.56	0.52	0.71	0.55	0.52	0.55	0.52	0.70	0.56	0.52	0.56	0.52	0.72
KNN	0.63	0.62	0.63	0.62	0.73	0.62	0.61	0.62	0.61	0.75	0.63	0.62	0.63	0.62	0.75	0.63	0.62	0.63	0.62	0.76	0.64	0.63	0.64	0.63	0.76
DT	0.65	0.64	0.65	0.64	0.86	0.64	0.63	0.64	0.62	0.83	0.67	0.65	0.67	0.66	0.85	0.66	0.65	0.66	0.65	0.87	0.66	0.66	0.66	0.65	0.87
NB	0.53	0.54	0.53	0.51	0.80	0.52	0.53	0.52	0.50	0.80	0.54	0.54	0.54	0.52	0.78	0.53	0.53	0.53	0.51	0.79	0.53	0.54	0.53	0.51	0.79
ADABOOST	0.61	0.57	0.61	0.58	0.84	0.62	0.57	0.62	0.58	0.83	0.62	0.59	0.62	0.59	0.83	0.62	0.59	0.62	0.59	0.83	0.62	0.58	0.62	0.58	0.84
ET	0.63	0.62	0.63	0.62	0.77	0.63	0.62	0.63	0.62	0.78	0.64	0.63	0.64	0.63	0.80	0.63	0.63	0.63	0.63	0.80	0.64	0.64	0.64	0.64	0.80
MLP	0.66	0.65	0.66	0.65	0.88	0.65	0.63	0.65	0.63	0.87	0.67	0.66	0.67	0.66	0.87	0.66	0.65	0.66	0.65	0.88	0.66	0.64	0.66	0.64	0.88
QDA	0.49	0.58	0.49	0.51	0.79	0.46	0.52	0.46	0.47	0.68	0.44	0.54	0.44	0.41	0.72	0.49	0.55	0.49	0.52	0.73	0.48	0.55	0.48	0.47	0.72
BAGGING	0.68	0.67	0.68	0.66	0.87	0.64	0.63	0.64	0.63	0.81	0.65	0.65	0.65	0.65	0.81	0.65	0.64	0.65	0.64	0.82	0.66	0.65	0.66	0.65	0.82

Table 7.5: Model performance across varying dataset(Classification)

Figure 7.16 shows the machine learning models' performance at various data sizes (20%, 40%, 60%, 80%, and 100%). Accuracy, precision, recall, and F1 scores are all comparatively constant for varying dataset sizes when using the Support Vector Machine (SVM), which performs steadily. The flat trendlines in Figure 7.16 serve as an example of this. The continuously high point in charts demonstrates how resilient algorithms like Random Forest (RF) and Gradient Boosting (GB) are at maintaining high-performance metrics over a range of dataset sizes. Figure 7.16 shows the mixed performance with Quadratic Discriminant Analysis (QDA) and Bagging with varying degrees of sensitivity to dataset size.



Figure 7.16: Line plot machine learning models' performance at various data sizes

In terms of the Area Under the Curve (AUC) scores, the evaluation of the models across different data sizes demonstrates some noteworthy differences. Gradient Boosting (GB) displays remarkable stability with AUC scores consistently close to 0.89, exemplifying their robust classification capabilities. Random Forest displays an upward trend in AUC scores, reaching approximately 0.90 at 80 and 100% dataset size. K Nearest Neighbour (KNN) model displays incremental improvements in AUC scores as the dataset size increases, underlining their adaptability to larger datasets

Multi-Layer Perceptron (MLP) consistently perform fine with AUC scores around 0.86-0.88 across all dataset sizes, emphasizing their reliability in capturing the true positive rate against the false positive rate. Quadratic Discriminant Analysis (QDA) displays varying degrees of sensitivity to data sizes, leading to fluctuations in AUC scores between 0.68-0.79. Figure 7.3 shows the AUC scores of the machine learning models' performance.

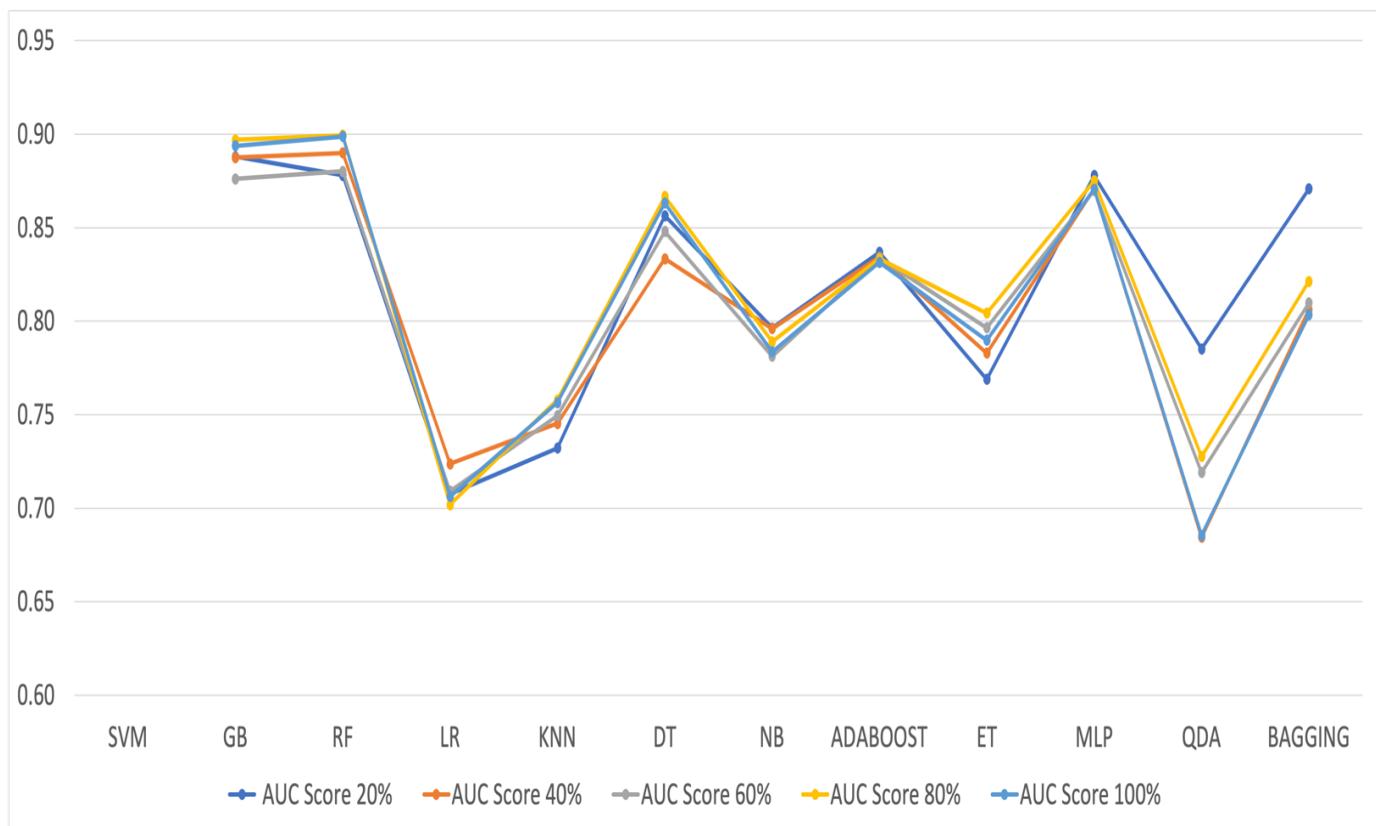


Figure 7.17: AUC scores of machine learning models' performance across various data sizes

7.8.10.2 Data Size Impact on Regression Model Performance

This section conducts a reliability analysis to investigate the effect of sample size on regression performance. Five different percentages (i.e., 20%, 40%, 60%, 80% and 100%) of data were utilized as training and testing sets to develop energy use rating prediction models. Twelve machine learning models were developed using five percentages of data. The most effective

predictive model is determined in the five cases of data availability using the model performance measures. In this chapter, the analysis of the result has shown major findings between the selected models as shown in Table 7.6 below. The basis of determining the best predictive model stipulates that model holding values closer to zero for MAE, MSE and RMSE are the good predictive model while values closer to one for R2 produced the best results. Support Vector Regression (SVR) demonstrates incremental improvement as the data size increases, reaching 0.46 at 100% data. Similarly, Artificial Neural Network (ANN) and Gradient Boosting (GB) reveal consistent improvements in R-squared values with larger datasets, showcasing their ability to leverage more data to make better predictions 0.73 and 0.75 respectively. Random Forest (RF) and Extra Trees (ET) consistently outperform other models, attaining a notable R-squared value of 0.78 at 100% data.

Likewise, other models such as KNN and DT among others. Contrarily, Linear Regression (LR), RIDGE and BR show relatively stable but moderate performance across different dataset sizes. All models showed increased or stable model performance across various data sizes, except LSR which shows a decrease in performance across the varying data sizes.

In relation to MAE, As the size of the dataset increases, SVR, ANN, GB, RF, and Bagging show consistently decreasing MAE values. This signifies their improved accuracy in predicting energy consumption with larger datasets. However, MAE values fluctuate across different dataset sizes for models like as LR, RIDE, LASSO, BR, and LSR, indicating sensitivity to changes in data quantity. In terms of RMSE, SVR, ANN, GB, RF, and Extra Trees (ET) consistently decreased RMSE values as the quantity of data size increased, highlighting their effectiveness in decreasing prediction errors with larger datasets.

in comparison with related work, Li et al., and Dong et al., applied SVM for predicting hourly load consumption on less than 5 instances with a result of 1.17(RMSE) and 0.99 (R2) respectively (Dong et al., 2005; C. Li et al., 2017). This result outperforms the performance of SVM in the research, though this can be the subject to the amount of instances used, based on the theoretical rationale proposed by a number of researchers that SVM is recognized for its generation of good results in small datasets (Aversa et al., 2016; Li et al., 2009b; Qiong Li et al., 2010). Furthermore, Dong et al., (Dong et al., 2005) applied SVM on a larger dataset of 507 instances with a result of 7.35 (RMSE), which performed significantly lower than the performance of SVM in this research using both 102,229 and 511,146 instances. Additionally, SVM shows improved performance across increasing datasets. The ANN produces good results and outperforms other studies as shown in Table 7.7 below. ANN still produces good results in both large and small datasets. Gradient boosting has not received much attention in this field

but performs remarkably well among others in terms of performance measures. GB presents promising potential for unprecedented results.

Table 7.2: Model performance across varying dataset(Regression)

REGRESSION	20% DATA			40% DATA			60% DATA			80% DATA			100% DATA		
MODEL	R-squared	MAE	RMSE												
SVR	0.13	113.33	155.28	0.16	113.32	152.77	0.18	113.41	154.52	0.17	113.97	157.12	0.19	111.58	153.98
ANN	0.73	65.46	89.53	0.73	63.46	86.53	0.77	59.77	82.52	0.80	55.23	77.78	0.81	52.55	74.37
GB	0.81	51.21	71.86	0.82	51.78	71.59	0.83	50.65	70.19	0.83	51.18	71.49	0.84	49.64	69.46
RF	0.79	55.69	76.87	0.78	56.98	78.64	0.81	52.96	73.90	0.80	54.02	76.62	0.82	51.80	73.26
LR	0.69	68.90	92.88	0.69	70.27	92.04	0.69	70.57	94.94	0.70	71.32	94.69	0.70	69.68	93.41
KNN	0.64	68.15	100.18	0.70	65.31	90.65	0.73	62.50	88.44	0.73	62.10	88.88	0.74	60.80	86.57
DT	0.63	71.72	100.68	0.65	71.05	99.07	0.69	66.81	94.48	0.71	66.18	93.11	0.72	63.61	90.65
ADABOOST	0.68	75.27	94.63	0.65	78.87	98.09	0.71	72.00	91.76	0.68	77.73	97.09	0.64	83.52	102.38
BAGGING	0.77	56.86	79.20	0.76	58.43	81.18	0.80	55.34	76.93	0.79	55.95	78.47	0.81	53.52	75.41
ET	0.73	60.82	85.64	0.74	60.34	84.53	0.78	56.66	79.86	0.78	57.57	81.32	0.79	55.41	78.89
RIDGE	0.69	68.90	92.88	0.69	70.27	92.04	0.69	70.61	94.93	0.70	71.36	94.69	0.70	69.67	93.41
LASSO	0.69	68.70	92.88	0.70	70.03	91.98	0.69	70.44	94.97	0.70	71.28	94.74	0.70	69.66	93.56
BR	0.69	68.88	92.88	0.69	70.25	92.04	0.69	70.60	94.93	0.70	71.35	94.69	0.70	69.66	93.41
LSR	-0.13	117.63	176.27	-0.17	118.55	180.33	-0.02	110.65	172.35	-0.15	119.71	184.85	-0.03	112.87	173.57
DNN	0.73	62.04	86.65	0.78	56.55	78.12	0.81	53.18	74.37	0.80	54.78	76.66	0.83	50.01	70.26

Figure 7.18 shows the RMSE line plot of different regression models performed at different dataset sizes (20%, 40%, 60%, 80%, and 100%). Support Vector Regression (SVR) demonstrates how RMSE values gradually decrease as dataset size increases, highlighting SVR's capacity to use more data to achieve better performance. A similar downward trend is shown in the Artificial Neural Network (ANN), which shows a steady improvement in prediction performance with more datasets. Deep Neural Network (DNN) continues to show a decrease in RMSE values, which supports its capacity to capture complex relationships within the data. Among these, Gradient Boosting (GB) is the most notable since it consistently shows declining RMSE values for all data sizes and has the lowest RMSE when 100% of the data is used. This pattern highlights GB's robustness in handling larger and more diverse datasets. Linear Regression (LR), while displaying stability in performance, has higher RMSE values compared to GB. AdaBoost displays some fluctuations in RMSE and R2 values, suggesting sensitivity to variations in dataset size. Extra Trees (ET) consistently exhibit a descending RMSE trend, signifying its resilience and adaptability across different data sizes. SVR and LSR produce the highest RMSE values of 153.98 and 173.57 respectively. Generally, the RMSE line plot (Figure 7.18) visually supports the quantitative analysis, delivering a clear representation of how data size impacts the predictive performance of regression models.

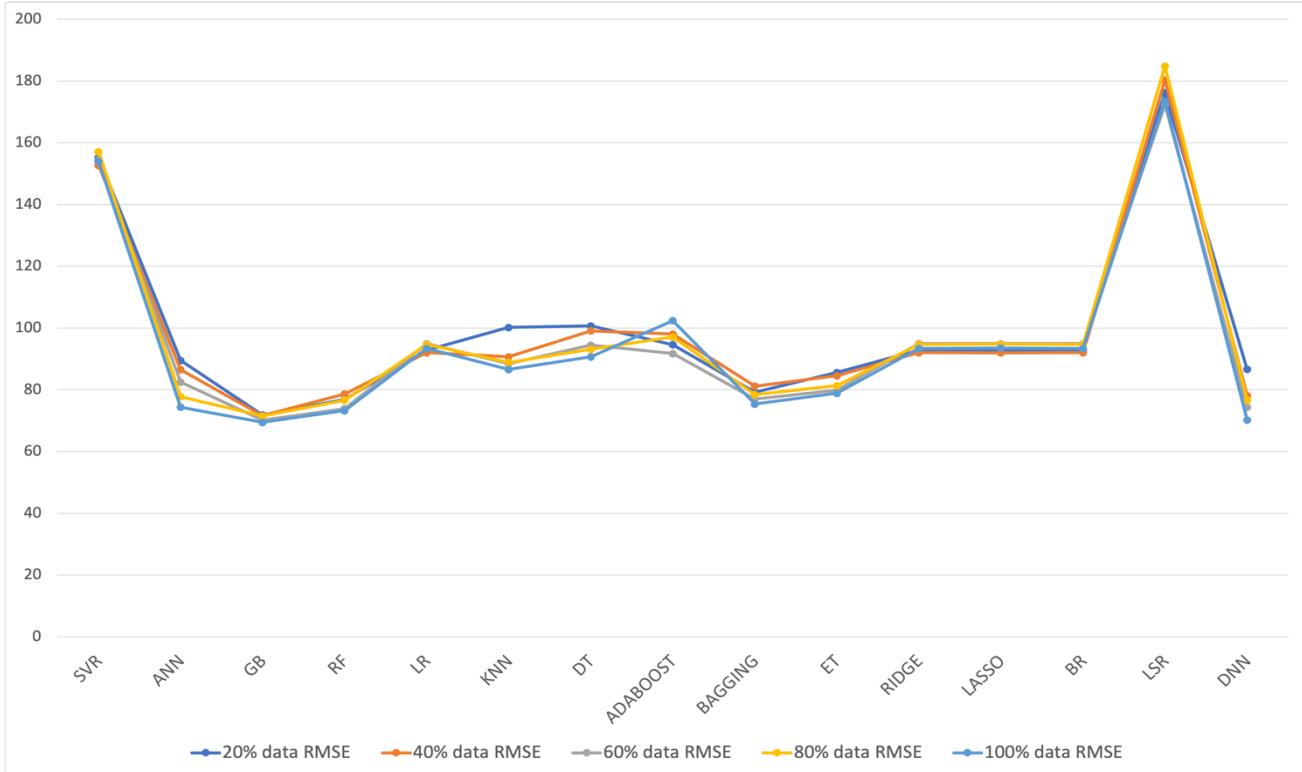


Figure 7.18: RMSE scores of machine learning models' performance across various data sizes

7.8.11 H8: Larger Data Size only Improve the Model Performance for certain ML Algorithms.

Similarly, the R² plot in Figure 7.19 shows ANN has a positive correlation between R-squared values and dataset size, signifying that the model's predictive power improves with larger datasets. KNN also exhibits a consistent upward trend in R-squared values, affirming its capacity to capture intricate patterns in the data. GB maintains a high and consistent R-squared value, demonstrating its resilience in identifying complex associations in the data.

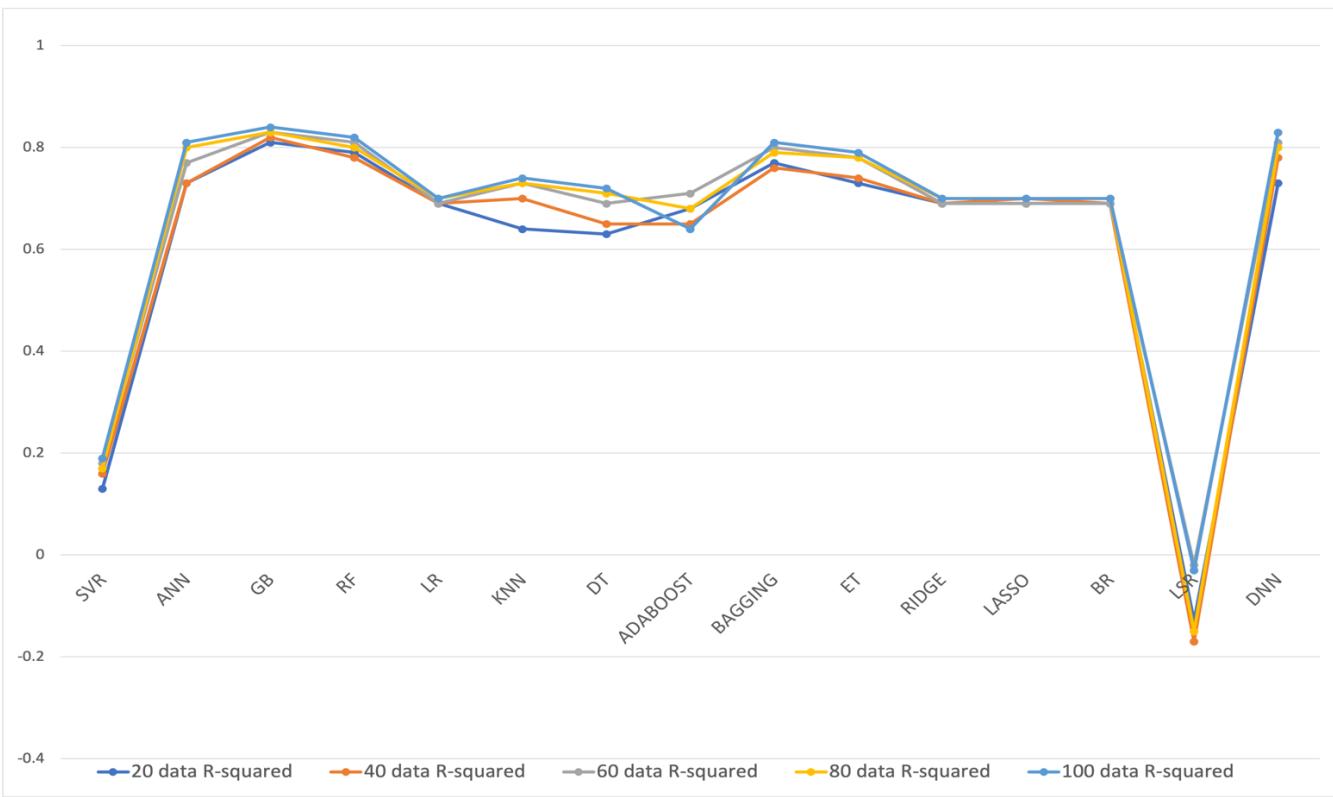


Figure 7.19:R-squared scores of machine learning models' performance across various data sizes

- **Null hypothesis H₀:** Larger data size only improves the model performance for certain ML algorithms.
- **Alternative hypothesis H_A:** Larger data size improves model performance for all ML algorithms.

The comprehensive analysis of regression models using data sizes of 20%, 40%, 60%, 80%, and 100% offers valuable insights on how data size affects predictive performance. The null hypothesis H₀ is supported by the observed patterns in R-squared and RMSE values, which indicate that some machine learning algorithms benefit from an increase in data size. For regression, one prominent example is Gradient Boosting (GB) which shows a steady and reliable increase in R-squared values with an increase in data size. The pattern in GB highlights its capacity to handle data more effectively, leading to better model fitting and accuracy. This aligns with the general hypothesis in machine learning that suggests predicting performance is positively impacted by larger datasets. ANN, KNN and DNN also display positive responses to increasing dataset sizes. Amber et al., (2018) particularly noted that DNN models are not as favourable in studies with a limited amount of data thus its performance relies heavily on a larger amount of data. These findings provide empirical evidence that larger datasets contribute to improved model performance for machine learning algorithms.

However, some studies such as (Mat Daut et al., 2017) have disputed that large data sizes do not generally lead to better performance and some ML algorithms such as Support Vector

Machine (SVM) thrive better in small datasets. In this research, it's important to note that not all ML algorithms exhibit improved performance with larger datasets. For instance, ML algorithms such as Ridge and Lasso showed no major improvements across the different data sizes, indicating that their performance may be influenced by factors beyond data size. It is also noted that lasso and ridge conduct feature selection implicitly during model development (Çiftsüren and Akkol, 2018). Therefore, this will further reduce the data size utilized to train the model. The variability in performance across algorithms highlights the nuanced relationship between data size and model effectiveness. Likewise, for classification, some ML algorithms showed negative effects with an increase in data size. For example, Bagging demonstrated a decline in performance with the increase in data size.

7.8.12 Big Data: Model Performance Analysis

The term 'Big Data' was introduced by John Mashey, who first utilized it in his presentation titled "Big Data and the Next Wave of InfraStress" (Diebold, 2012). The definition of Big Data is complex due to the relative nature of the term 'big,' but the concept of big data primarily revolves around three main characteristics, prominently known as the three V's: velocity, volume, and variety (Ivanov et al., 2013). The well-known 3V architectural paradigm for Big Data was introduced by (Laney, 2001). Volume denotes the data's size, velocity relates to the speed of data generation and the need for its analysis, and variety represents the extent of data variability (Kitchin and McArdle, 2016).

The statistical and machine learning models were developed in the Amazon Web Services (AWS) environment, employing key services tailored for machine learning tasks. The selected data server for this solution is Amazon SageMaker, a fully managed service devised for the end-to-end lifecycle of building, training, and deploying machine learning models at scale. This platform streamlines the process of machine learning model development and offers a scalable infrastructure for handling large datasets.

In terms of computing resources, the 'Instance' option chosen for the development environment was 'm4.large.' This choice was driven by considerations of cost, selecting an instance type that balances computational power with affordability. The data containing the building metadata and building energy consumption values was stored in an Amazon S3 bucket. S3, or Simple Storage Service, is an object storage service that enables the storage and retrieval of large amounts of data. By employing S3, the system is shaped to handle the storage requirements of the data involved in the Big Data Analytics development of BE-CPM (Building Energy Consumption Prediction Model). This AWS environment provides a robust foundation for developing and deploying machine learning models at scale, ensuring efficient management of

data, cost-effectiveness, and scalability, which are vital aspects of statistical and machine learning model development. Table 7.8 shows the model performance using big data analytics.

Table 7.3: Big-data model performance

S/N	REGRESSION					CLASSIFICATION				
	Model	Time(sec)	R-squared	MAE	RMSE	Time(sec)	Accuracy	Precision	Recall	F1
1.	BD-SVM	37825.76	0.19	111.58	153.98	34781.76	0.59	0.57	0.59	0.57
2.	BD-GB	44.177	0.84	49.64	69.46	382.91	0.69	0.68	0.69	0.68
3.	BD-RF	145.331	0.82	51.8	73.26	42.469	0.69	0.68	0.69	0.68
4.	BD-LR	0.1782	0.7	69.68	93.41	10.97	0.55	0.52	0.55	0.52
5.	BD-KNN	1.1688	0.74	60.8	86.57	0.813	0.64	0.63	0.64	0.63
6.	BD-DT	1.2565	0.72	63.61	90.65	0.473	0.65	0.64	0.65	0.64
7.	BD-ADABOOST	41.9755	0.64	83.52	102.38	25.74	0.62	0.58	0.62	0.59
8.	BD-ET	123.936	0.79	55.41	78.89	35.100	0.63	0.62	0.63	0.62
9.	BD-MLP	30474.76	0.81	52.55	74.37	385.889	0.65	0.63	0.65	0.63
10.	BD-BAGGING	146.5838	0.81	53.52	75.41	38.926	0.69	0.68	0.69	0.68
11.	BD-RIDGE	0.37316	0.7	69.67	93.41	—	—	—	—	—
12.	BD-LASSO	0.10609	0.7	69.66	93.56	—	—	—	—	—
13.	BD-BR	0.2365	0.7	69.66	93.41	—	—	—	—	—
14.	BD-LSR	930.378	-0.03	112.87	173.57	—	—	—	—	—
15.	BD-DNN	5715.613	0.83	50.01	70.26	—	—	—	—	—
16.	BD-QDA	—	—	—	—	0.3461	0.49	0.47	0.49	0.47

Using Big data analytics, the results produced from the classification and regression models offer insightful information on the effectiveness and speed of statistical/machine learning tools. The time consumed for model training, as represented in seconds, indicates distinct differences among the models. The computational efficiency of these models differs greatly in regression. The remarkably short processing times of LASSO (BD-LASSO) and Linear Regression (BD-LR) show that these statistical techniques may evaluate energy use more quickly. On the other hand, models such as Support Vector Machine (BD-SVM), Multi-layer Perceptron (BD-MLP) and Deep Neural Network (DNN) require substantially more time, underlining potential drawbacks in terms of speed.

7.8.13 H9: Statistical/ML Tools can Assess Energy Consumption Faster than the Traditional Method

For classification, Naïve Bayes (BD-NB) and decision tree (BD-DT) demonstrate rapid processing). Additionally effective are Random Forest (BD-RF) and Bagging (BD-BAGGING), which demonstrate the model's capacity to strike a balance between speed and accuracy. Given their longer execution times, Support Vector Machine (BD-SVM) and Multi-layer Perceptron (BD-MLP) may not be the most time-efficient models for classification tasks. The null hypothesis that statistical and machine learning techniques are more efficient than traditional computer-based energy simulation tools is investigated.

- **Null hypothesis H_0 :** Statistical/ML tools can predict energy consumption faster than the traditional method.
- **Alternative hypothesis H_A :** Statistical/ML tools cannot predict energy consumption faster than the traditional method.

When comparing the processing times of different machine learning models to more conventional computer-based energy simulation models such as EnergyPlus and DOE2, there is a noticeable speed advantage, especially in regression and classification tasks for energy consumption prediction. These machine learning models such as Linear Regression (BD-LR), LASSO (BD-LASSO), Gradient Boosting (BD-GB), and Random Forest (BD-RF) in particular—have significantly reduced processing times, which attests to their effectiveness. In various fields such as healthcare (Jin et al., 2006; Leyh-Bannurah et al., 2018; Zheng et al., 2019), pollution prediction (Balogun et al., 2021; Sulaimon et al., 2021), and bankruptcy prediction(Alaka et al., 2018; Barboza et al., 2017; N. Wang, 2017), among others, statistical and machine learning algorithms are producing good performance in terms of accuracy and computational efficiency.

EnergyPlus and DOE2, two well-known computer-based energy simulation models, are highly accurate and precise at modelling the energy performance of buildings. But since these simulation engines are so complicated, they can take a long time to execute—up to an hour or more in the case of huge, complex building models. Yu et al., (2015) conducted an investigation on the impact of the envelope on energy consumption using Energy Plus. It was emphatically stipulated that the user interface of EnergyPlus is extremely unfriendly, time-consuming and offers poor visibility. This prolonged duration can be a bottleneck in scenarios where a quick assessment of energy consumption is required, impeding the ability to make decisions in real-time or near real-time. Conversely, machine learning models such as Extra Trees (BD-ET) and

Random Forest(BD-RF) show noticeably faster processing times without sacrificing prediction accuracy.

Machine learning models are particularly appealing for applications that require immediate insights or quick assessments due to their shorter processing times. They offer a strong substitute for simulation models that require more computation, such as DOE2 and EnergyPlus(Yu et al., 2015). The trade-off between speed and computational complexity demonstrates machine learning's adaptability in situations when striking a balance between accuracy and efficiency is essential. Thus, the null hypothesis holds true, indicating that specific statistical and machine-learning techniques may effectively predict energy consumption faster than traditional computer-based energy simulation models.

Although the GB model achieved the highest accuracy across all models using 10 features, it does not outperform GB and other models' performance using over 13 features. However, the disparity in these values is by such a small margin that it can be considered statistically insignificant. Nevertheless, to further improve the performance of the models using 10 features, it is necessary to explore hyperparameter tuning of the algorithm. Considering the highest accuracy was produced using the 10 most relevant features from chi-square, those features will be used for subsequent model development and optimization.

7.8.14 Hyperparameter Tuning for Model Optimization

This involves the process of fine-tuning the parameters of statistical and machine learning models to enhance their performance. In both classification and regression tasks, choosing the optimal hyperparameters is crucial for achieving high accuracy and predictive power. Proper hyperparameter tuning can significantly improve model performance by preventing overfitting, reducing bias, and ensuring better generalization to new data.

7.8.14.1 Classification Model Optimization – Hyperparameter Tuning

To satisfy the third objective of this research, Hyperparameter tuning did help produce better performance in certain algorithms such as RF, DT and KNN. Hyperparameter tuning does improve the accuracy of models according to Singh et al., (2021). However, it must be noted that it is computationally expensive as it takes longer training and testing time. the correlation of parameter tuning from the perspective of the level of increase in performance and the amount of computational cost increased. However, the overall time taken is still less than the traditional simulation method.

The optimization process involved hyperparameter tuning through the application of grid search and cross-validation which helps to fine-tune the model to better capture underlying patterns within the data. Grid search was utilized to systematically search across a predefined set of hyperparameter permutations and the model's performance was evaluated using cross-validation to determine the best combination that engenders the best performance based on the chosen performance evaluation metric. The optimal parameters selected using grid search for each model are as follows:

1. **SVM (Support Vector Machine):** (kernel='rbf', probability=True, degree=3, gamma='scale')
2. **Gradient Boosting:** (learning_rate=0.1, max_depth=5, min_samples_leaf=1, min_samples_split=5, n_estimators=100, subsample=1.0, random_state=123)
3. **Random Forest:** (max_depth=None, min_samples_leaf=2, min_samples_split=5, n_estimators=100, random_state=123)
4. **Logistic Regression:** (C=1, penalty='l2', random_state=123)
5. **K-Nearest Neighbors:** (n_neighbors=5, p=1, weights='uniform', algorithm='auto', leaf_size=30)
6. **Decision Tree:** (max_depth=10, max_features='auto', min_samples_leaf=2, min_samples_split=2, criterion='gini', splitter='best')
7. **Naive Bayes:** No specific hyperparameters for Naive Bayes as it is a probabilistic model.
8. **AdaBoost:** (learning_rate=0.1, n_estimators=100)
9. **Extra Trees:** (n_estimators=100, max_depth=None, min_samples_split=2, criterion='gini', random_state=42)
10. **Multi-layer Perceptron (Neural Network):** (activation='relu', solver='adam', alpha=0.0001)
11. **Quadratic Discriminant Analysis:** No specific hyperparameters for Quadratic Discriminant Analysis.
12. **Bagging Classifier:** (base_estimator=None, n_estimators=50, random_state=75)

Following the model optimization, there is a notable improvement in the performance values of various machine learning models. The most prominent improvements are detected in the decision tree, Random Forest, and Bagging models which exhibit increased accuracy, precision, recall, and F1 scores. For example, RF achieved 0.64 before optimization and 0.69 after model optimization. These improvements accentuate the effectiveness of the model optimization in amplifying the models' predictive capabilities. Conspicuously, models such as SVM, Logistic Regression, Gradient boosting and Naive Bayes show more modest improvements, indicating that the optimization process had a varying impact across varying algorithms. Overall, the optimization effort resulted in better models. Gradient Boosting (GB) shows an increase in accuracy from 0.68 to 0.69 which is at par with the same level of accuracy using over 13 features. The performance values obtained using chi-square over 13 features and the optimized models using only 10 features further highlight the importance of meticulous

feature engineering and selection processes, as the presence of excessive or irrelevant features can lead to increased model complexity without essentially improving predictive performance. Table 7.9 below shows the difference in performance before and after model optimization.

Table 7.4: Model performance before and after optimization (Classification)

Model	Before optimization				After Optimization			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SVM	0.58	0.54	0.58	0.56	0.59	0.55	0.59	0.57
GB	0.68	0.67	0.68	0.67	0.69	0.68	0.69	0.68
RF	0.64	0.63	0.64	0.63	0.69	0.68	0.69	0.68
LR	0.55	0.50	0.55	0.52	0.55	0.51	0.55	0.51
KNN	0.58	0.57	0.58	0.57	0.62	0.61	0.62	0.61
DT	0.59	0.59	0.59	0.59	0.65	0.64	0.65	0.64
NB	0.53	0.54	0.53	0.52	0.54	0.54	0.54	0.52
ADABOOST	0.55	0.56	0.55	0.53	0.62	0.59	0.62	0.59
ET	0.62	0.61	0.62	0.62	0.63	0.62	0.63	0.62
MLP	0.65	0.64	0.65	0.64	0.65	0.64	0.65	0.64
QDA	0.49	0.57	0.49	0.46	0.48	0.52	0.48	0.49
BAGGING	0.62	0.61	0.62	0.62	0.65	0.64	0.65	0.64
VOTING	0.70	0.69	0.70	0.68	0.70	0.69	0.70	0.68

The Receiver Operating Characteristic (ROC) curves offer valuable insights into the classification performance of the models across multiple classes. It displays the ability of the models to differentiate between positive and negative instances. Figure 7.20 displays the ROC area under the curve for SVC, GB, RF and LR. Class A to G in these figures represent the target variable (Energy rating)

For SVC, it displays good classification in Class C ($AUC = 0.80$) and Class G ($AUC = 0.82$), denoting strong performance in distinguishing positive and negative instances for these classes. However, it struggles significantly in Class B ($AUC = 0.44$), where the AUC is relatively low. GB exhibits strong classification performance with consistently high AUC values across all classes. Particularly exceptional are the performances in Class F ($AUC = 0.96$) and Class G ($AUC = 0.98$), denoting the model's effectiveness in classifying instances for these classes. Likewise, RF produced good performance across all the classes. Logistic regression however produced varying performance values across the classes. While it delivers good performance in Class F ($AUC = 0.81$) and Class G ($AUC = 0.87$), it struggles to produce good performance in Class D ($AUC = 0.64$), where the AUC is notably lower, reflecting slight difficulties in classifying instances for this class.

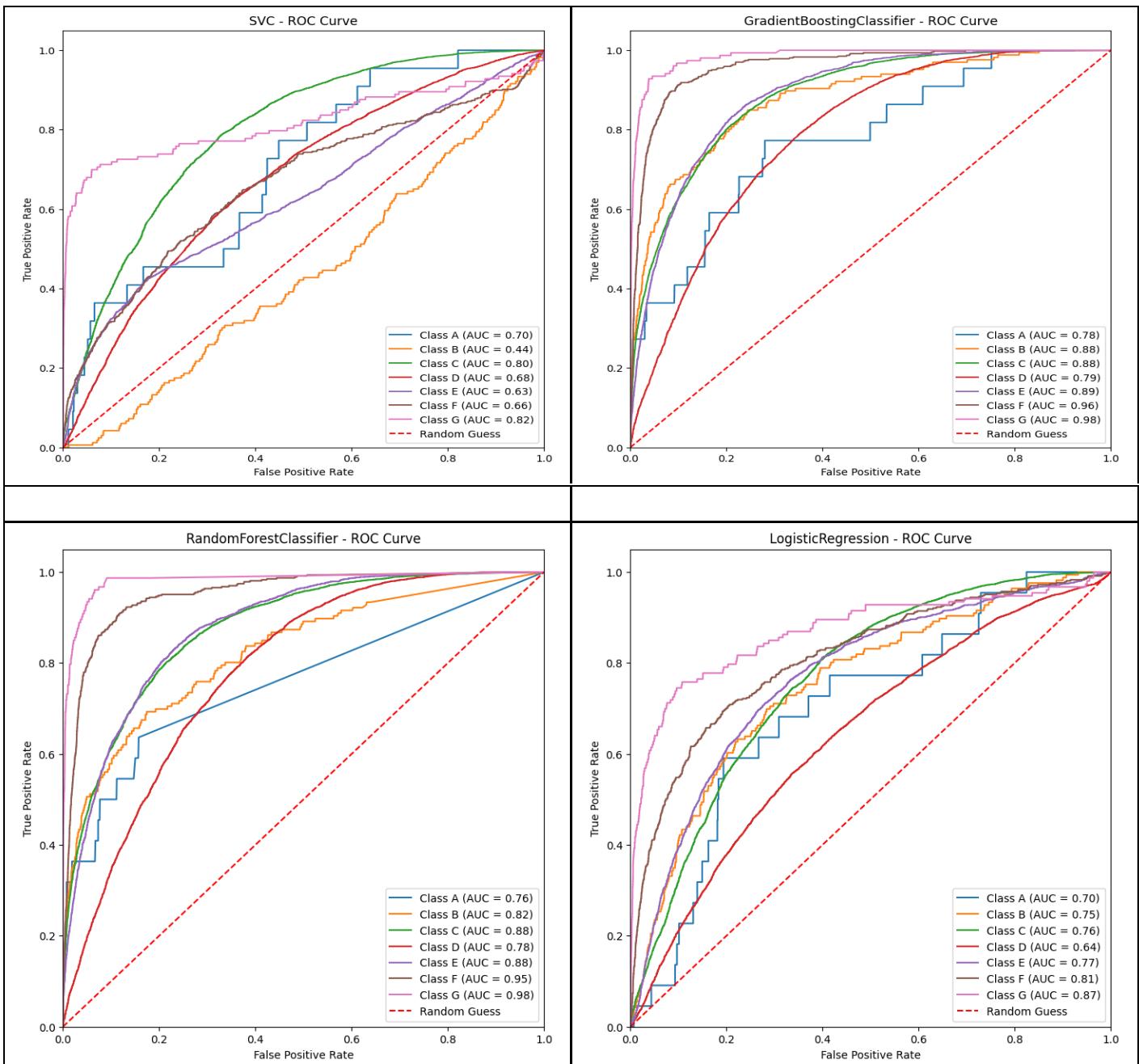


Figure 7.20: ROC curve for SVC, GB, RF, LSR

Figure 7.21 shows ROC AUC values for KNN, DT, AdaBoost, and GaussianNB models. It shows their performance across different classes. For KNN, the AUC values range from 0.53 for Class A to 0.81 for Class C. This indicates that KNN performs moderately well, with varying performance values across different classes. DT performed relatively better, with AUC values ranging from 0.71 for Class A to 0.90 for Class F.

Likewise, AdaBoost produced even higher AUC scores for Class F (0.94) and Class G (0.98). These suggest that AdaBoost is very effective in classifying instances for most of the classes.

On the other hand, GaussianNB demonstrates moderate performance, with AUC values ranging from 0.66 to 0.91 for Class D and Class G respectively.

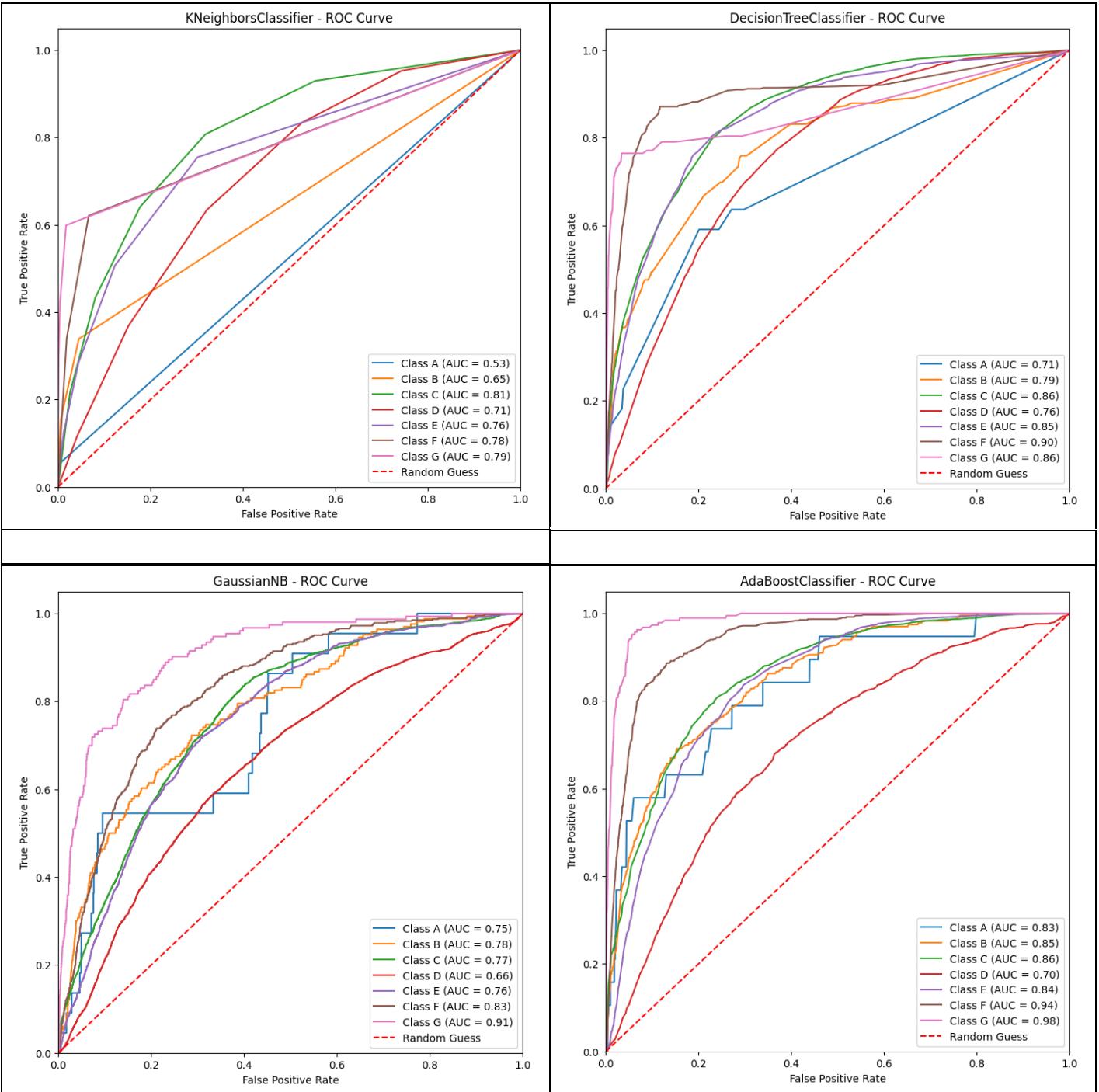


Figure 7.21: ROC curve for KNN, DT, NB, Adaboost

Figure 7.22 shows the ROC AUC values for Extra Trees (ET), Multi-layer Perceptron (MLP), Quadratic Discriminant Analysis (QDA), and Bagging models. MLP exhibits strong performance with AUC values ranging from 0.71 to 0.94 for Class A and Class G respectively. The overall performance of MLP shows effectiveness in classifying specific classes. Contrarily, QDA elicit low performance for class A and class F with AUC scores of 0.54 and 0.50 respectively. Bagging exhibits good overall performance, with AUC values ranging from

0.63 to 0.94 for Class A and Class G respectively. ET also performs well across different classes, with AUC values ranging from 0.61 for Class A to 0.95 for Class G.

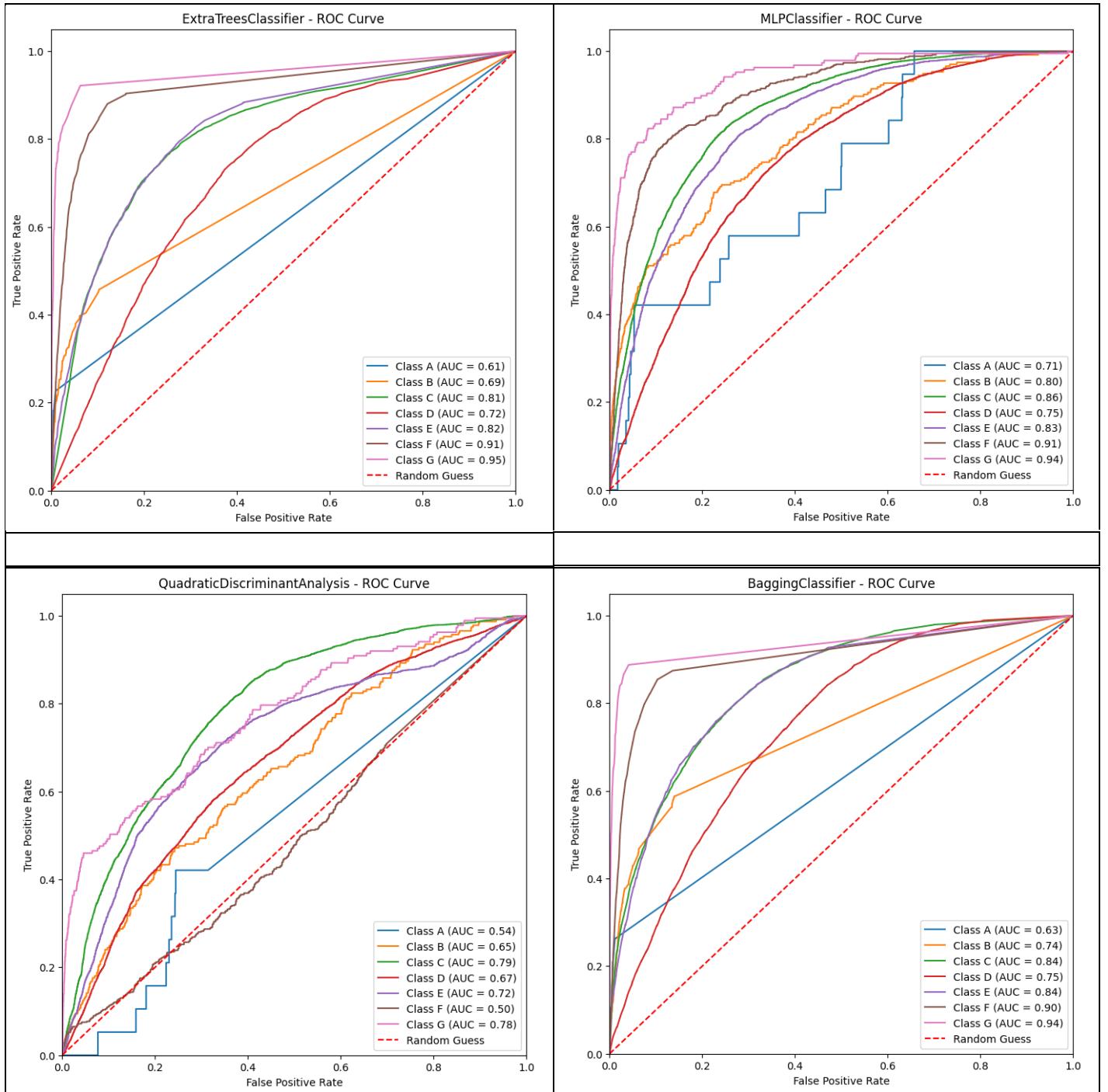


Figure 7.22: ROC curve for ET, MLP, QDA, Bagging

The ROC curve analysis demonstrates that all models have their individual strengths and weaknesses across different classes. However, Gradient Boosting consistently emerged as the top-performing model across multiple classes. The optimal selection of a model might differ based on the specific complexities of each class.

To summarize the overall performance of the various models for each class based on their AUC values:

1. Class A:

- Best Model: Adaboost(AUC = 0.83)
- Other Strong Performers: Gradient Boosting (AUC = 0.78), Random Forest (AUC = 0.76)

2. Class B:

- Best Model: Gradient Boosting (AUC = 0.88)
- Other Strong Performers: Adaboost (AUC = 0.85), Random Forest (AUC = 0.82)

3. Class C:

- Best Model: Gradient Boosting (AUC = 0.88)
- Other Strong Performers: Random Forest (AUC = 0.86), Extra Trees (AUC = 0.86)

4. Class D:

- Best Model: Gradient Boosting (AUC = 0.79)
- Other Strong Performers: Random Forest (AUC = 0.78), Decision Tree (AUC = 0.76)

5. Class E:

- Best Model: Gradient Boosting (AUC = 0.89)
- Other Strong Performers: Random Forest (AUC = 0.88), Decision Tree (AUC = 0.85)

6. Class F:

- Best Model: Gradient Boosting (AUC = 0.96)
- Other Strong Performers: Random Forest (AUC = 0.95), Adaboost (AUC = 0.94)

7. Class G:

- Best Model: Gradient Boosting (AUC = 0.98)
- Other Strong Performers: Random Forest (AUC = 0.98), Adaboost (AUC = 0.98)

Gradient boosting and Random Forest emerged as the best or second-best models in classifying all the classes in the dataset. This result indicates the effectiveness of ensemble-based algorithms in handling complex patterns present in the data. Gradient Boosting conducts an iterative boosting process that centres on correcting the errors of previous models which exhibited better predictive performance. Random Forest, with its ensemble of decision trees and inherent capacity to handle various types of data, also appeared to be a strong contender.

Given the strengths of these models, a tactical approach was initiated to leverage on their individual capabilities. A Voting Classifier, an ensemble technique that implements a combination of the predictions of multiple base predictors or estimators, was employed to develop a stronger and more robust model(Tsai et al., 2014). By capitalizing on the strengths of Gradient Boosting and Random Forest, the Voting Classifier will potentially capture various patterns of the data and produce a more comprehensive and accurate prediction across all classes. This ensemble method not only improves the predictive performance but also provides a more reliable and stable model by alleviating the individual weaknesses of each algorithm. The choice to combine these two top-performing models is a strategic approach aimed at achieving better classification results across the different target classes.

Figure 7.23 shows the accuracy value and ROC AUC values for the voting classifier. The Voting Classifier achieved a performance accuracy value of 70% which outperforms the accuracy values using the models unilaterally. The Voting Classifier produces very good performance across all the classes, ranging from 0.79 to 1.0 for Class B and Class G respectively.

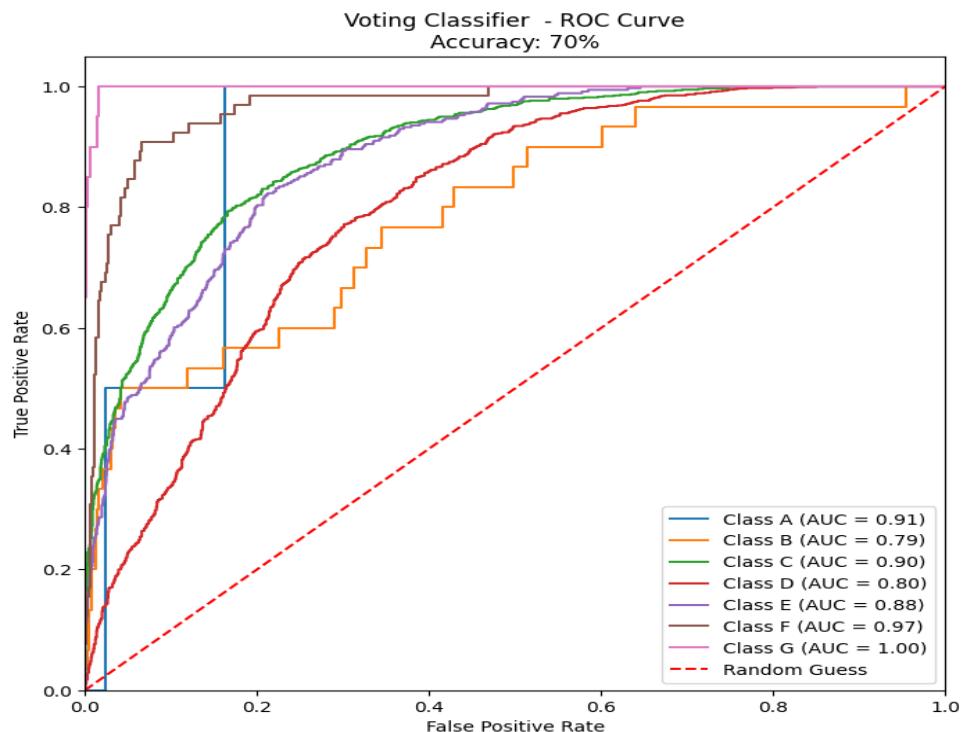


Figure 7.23: ROC curve for voting

7.8.14.2 Regression Model Optimization – Hyperparameter Tuning

For regression, to further improve the performance of the models using 10 features, hyperparameter tuning was also explored in the algorithm. Considering GB and RF achieved the highest r-squared value using the 10 most relevant features from the random forest, the aforementioned features will be utilized for subsequent model development and optimization. The optimal parameters selected using grid search for each model are as follows:

1. **SVM (Support Vector Machine):** (kernel='rbf')
2. **Multi-layer Perceptron (Neural Network):** (activation='relu', solver='adam', alpha=0.0001)
3. **Gradient Boosting:** (random_state=123)
4. **Random Forest:** (max_depth=20, min_samples_split=10, n_estimators=100, random_state=123)
5. **Linear Regression: ()**
6. **K-Nearest Neighbors:** (n_neighbors=7, weights='uniform')
7. **Decision Tree:** (max_depth=10, min_samples_split=10)
8. **AdaBoost:** (learning_rate=0.1, n_estimators=50)
9. **Extra Trees:** (max_depth=20, min_samples_split=10, n_estimators=200, random_state=123)
10. **Ridge:** (alpha=2.0)
11. **Lasso:** (alpha=2.0)
12. **BayesianRidge:** Default
13. **LogisticRegression:** Default
14. **Bagging Classifier:** (max_features=1.0, max_samples=0.5, n_estimators=200)

After model optimization through hyperparameter tuning and cross-validation, there is a notable improvement in the performance values of specific machine learning models namely SVR, DT and Adaboost. The most noticeable improvements appeared in the decision tree, and Adaboost models which exhibit significant increase in R-square value. For example, DT achieved 0.72 before optimization and 0.81 after model optimization. Adaboost also showed an increase in r-square value from 0.64 to 0.71. Overall, the optimization efforts resulted in better models. Table 7.9 below shows the difference in performance before and after model optimization.

Table 7.5:Model performance before and after optimization (Regression)

Model	Before optimization			After Optimization		
	R-squared	MAE	RMSE	R-squared	MAE	RMSE
SVR	0.19	111.58	153.98	0.69	68.74	95.09
ANN	0.81	52.55	74.37	0.80	54.28	76.81
GB	0.84	49.64	69.46	0.85	48.21	67.14
RF	0.82	51.80	73.26	0.83	49.51	69.53
LR	0.70	69.68	93.41	0.70	69.68	93.41
KNN	0.74	60.80	86.57	0.75	59.86	86.03
DT	0.72	63.61	90.65	0.81	52.51	74.00
ADABOOST	0.64	83.52	102.38	0.71	70.55	92.71

BAGGING	0.81	53.52	75.41	0.83	49.87	70.33
ET	0.79	55.41	78.89	0.83	49.28	69.65
RIDGE	0.70	69.67	93.41	0.70	69.67	93.41
LASSO	0.70	69.66	93.56	0.70	69.65	93.71
BR	0.70	69.66	93.41	0.70	69.66	93.41
LSR	-0.03	112.87	173.57	-0.03	112.87	173.57
DNN	0.83	50.01	70.26	0.83	50.01	70.26

The top-performing models appear to be RF, GB, and DNN. These models consistently display high R-squared values and relatively low root mean squared error (RMSE) across various feature selection methods and varying numbers of features. Particularly the optimized Gradient Boosting model stands out with an R-squared of 0.85, demonstrating its robustness in capturing complex connections within the data. Likewise, Random Forest and Deep Neural Network also consistently perform well. Figure 7.24 shows a pie chart of the performance of all models after optimization (Both classification and Regression).

In both the classification and regression tasks for predicting energy consumption, various machine learning models were evaluated based on their performance using Accuracy and r2 squared respectively. In classification, where the energy performance rating served as the dependent variable, the models exhibited varying degrees of accuracy. Notably, Gradient Boosting (GB), Random Forest (RF), and the ensemble method Voting Classifier (GB and RF) emerged as the top performers, achieving accuracies of 0.69, 0.69 and 0.70, respectively. These models exhibited robust classification capabilities, providing some accurate predictions of energy performance ratings.

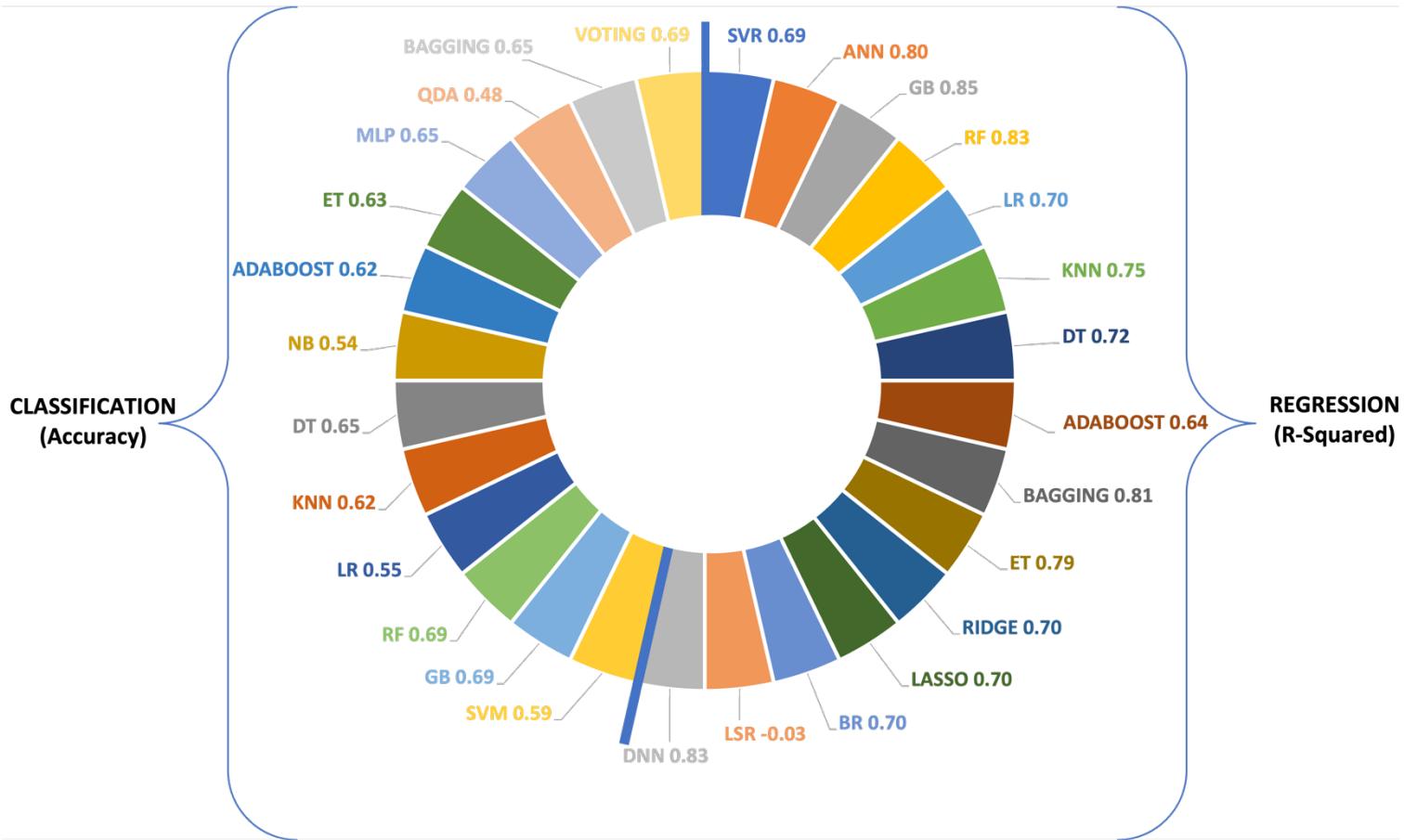


Figure 7.24: Model performance for both classification and regression

Similarly, in the regression task, where the dependent variable was the predicted energy consumption values, the models were evaluated using the coefficient of determination (R-squared). The best methods were found to be Gradient Boosting (GB), Random Forest (RF), and Deep Neural Network, with r² values of 0.85, 0.83, and 0.83, respectively. These models showed strong predictive capabilities, capturing a significant quantity of the variance in energy consumption values and demonstrating their efficacy in regression tasks. Overall, Gradient Boosting and Random Forest had high performance in both classification and regression analyses, indicating their adaptability and efficacy in different predictive tasks related to energy consumption.

7.9 Chapter Implications and Key Insights

This research presents the theoretical and practical implications of this research. This research presents both the theoretical and practical implications, highlighting the contribution to the existing body of knowledge and its application in real-world scenarios.

7.9.1 Theoretical Implications

This research explored the utilization of various FS techniques in developing several machine learning models to conduct a clear and unbiased comparison. Some studies suggest that feature selection is essential for the optimum performance of the model (Alaka et al., 2018; Balogun et al., 2021; Zhang and Wen, 2019a; Hai-xiang Zhao and Magoulès, 2012a); some studies suggest that feature selection is unfavourable to certain algorithms (Alaka et al., 2019, 2018; Olu-Ajayi et al., 2022a). However, this research shows that feature selection can have a positive or negative impact on the predictive accuracy, depending on the algorithm selected. This supports previous research, which states that the achievement of a good predictive accuracy of a model is highly predicated on the type of algorithm and feature selection method chosen (Olu-Ajayi et al., 2022b). Therefore, if an appropriate feature selection method is not selected for the specific ML algorithm used, feature selection will not result in good accuracy. For example, SVC produced good results for Mutual Information and ANOVA, which are both filter methods. This could suggest that SVC is best suited to filter methods. However, some models achieved better accuracy without feature selection such as Gradient Boosting (GB) and Adaboost among others.

One general hypothesis in the machine learning world, states that the larger the data utilized for model training, the better the performance (Dalal, 2018; Goyal et al., 2020; Kaur and Gupta, 2017; Lee et al., 2011). In this research, a reliability analysis was conducted to substantiate or disprove this hypothesis. Different machine learning models were developed using five sample sizes (20%, 40%, 60%, 80%, 100%). This finding engendered from the analysis suggests that the larger the data does not always lead to a better result. Although, notably, a single study is not enough to substantiate this conclusion, and this should be subject to further investigation. Some studies have corroborated certain conclusions, for example, Deep Neural Networks (DNN) showed a relative increase in predictive accuracy based on the increase in data size. Past studies have shown that large data has an effect on certain machine learning algorithms such as neural networks, (Amasyali and El-Gohary, 2018; Bourhnane et al., 2020). Also, the study by Bourhnane et al., 2020 stipulates that neural network algorithms are dominant with big datasets, as they require sufficient data to train the model. On the other hand, SVC shows no significant increase in predictive accuracy based on the increase in the data size. This can be subject to the conclusion that Support Vector Machines (SVM) are recognized for their ability to deliver good results effectively regardless of data size (Li et al., 2009; Qiong Li et al., 2010). Therefore, the general hypothesis on large data and better predictive performance is only peculiar to certain machine learning algorithms.

7.9.2 Practical Implications

The identification of the most relevant features that influence the energy performance of a building is important for several reasons and at various stages. During the development of a building energy prediction model, the utilization of only the relevant variables can improve the accuracy of the model. This can also help in reducing the model's complexity and avoiding overfitting. Additionally, this could reduce the time-consuming effort and high cost required to collect data on all variables that may affect building energy consumption. Furthermore, the use of only the relevant variables for an energy prediction model can reduce the computational cost. Moreover, identifying of the most relevant features is also imperative at the design stage of a building because this enables the building designer to discern which building features require optimization to achieve a low potential energy consumption outcome.

Overall, the choice of the most suitable feature selection method can help in identifying the most relevant features that contribute to the energy use of a building. This research identifies the most suitable FS method for specific ML algorithms, for instance, SVC, ET, MLP and bagging were found to be most suited to filter feature selection methods such as chi-square, mutual information, and ANOVA, among others. In terms of technical, social and economic implications, technical expertise is required to implement different feature selection methods and exploring various FS methods to identify the most suitable for an ML algorithm it is more labour-intensive. This research delivers the most suitable FS method for certain ML algorithms, reducing the time-consuming process of exploring various FS methods.

In real-world situations, the process of collecting and utilizing data could raise privacy concerns, particularly with personal data such as occupancy details among others. The data of numerous features are often collected. However, in most cases, only a few of these features may be related to the target output (Kira and Rendell, 1992). The identification of the most relevant feature will help limit the types of data required and reduce the cost of data collection. Developing a high-performing building energy consumption prediction model can help organizations optimize the use of energy resources, leading to cost savings and improved sustainability. Additionally, the implementation of feature selection and reliability analysis in this research can help model developers identify the ML algorithms that are sensitive to changes in the data or feature size.

7.10 Chapter Summary

This chapter conducts the exploration and application of AI/ML techniques in predicting and optimizing building energy consumption. It delivers the methodology to convey in detail the approach employed in this chapter. This chapter conducts model development, with a focus on both classification and regression. Several hypotheses regarding the impact of feature selection on machine learning performance and the role of weather data are examined. Additionally, the potential advantage of deep learning algorithms over classical machine learning algorithms is explored. Subsequently, this chapter addresses the hypotheses regarding the impact of large data size on model performance and the efficacy of statistical/ML tools in assessing energy consumption. The section "Big data for Energy prediction" examines the hypothesis regarding the influence of data size on model performance. The Classification and "Regression Models were further optimized to achieve predictive models. Through these investigations, the chapter contributes to advancing the understanding and application of AI/ML in optimizing building energy consumption. Lastly, this chapter also discusses the implications of this chapter, both theoretical and practical, highlighting its contributions to the applications in the field of building energy consumption prediction.

CHAPTER EIGHT

8.0 REVERSE ENGINEERED SYSTEM AND VALIDATED FRAMEWORK FOR BUILDING ENERGY PREDICTION

8.1 Chapter Overview

This chapter delivers the reverse engineered systems for energy assessment at the design stage which offers a back-to-front approach for building designers. This will avail building designer the ability to simply specify the desired or target energy value and receive optimal value for each building feature to achieve the target energy value. Additionally, the chapter provides a comprehensive validated framework for energy consumption prediction based on the investigation and outcomes of experiment conducted in chapter 6 and 7 of this research. The insights extrapolated from empirical experiments led to the creation of a validated framework. The existence of a validated framework for tool selection across different criteria such as variable selection, data size, error rate, and computational cost, will aid more informed decisions in tool selection for building energy consumption prediction, potentially leading to more accurate models. This will potentially equip researchers and practitioners with a streamlined framework for tool selection, thereby saving time and effort consumed on exhaustive comparative analysis of several tools to identify the most suitable for a specific situation.

8.2 Validated Framework

The development of an accurate and reliable Building Energy consumption Prediction model is considered a multifaceted task that demands a strategic selection of tools for prediction. The most prominent methods of tool selection are the selection of tools based on prevalence in research while some are arbitrary. This often leads to a high consumption of time and effort in exhaustive comparative analysis of several tools to identify the most suitable for a specific situation. Given that there is no one-size-fits-all all AI-based tool, comparison and evaluation of these tools is paramount to avail BEP model developers a guideline towards an informed selection of tools (K. Amasyali and El-Gohary, 2021). Ideally, model developers should recognise the benefits and drawbacks of the existing tools to achieve good outcomes in relation to certain criteria (e.g., data size, error rate etc.). This will ascertain the appropriate tool is

applied in the appropriate situation for the appropriate data features and purpose. It is proffered that well-informed tool selection for a certain condition can result in the development of more accurate prediction models and more efficient comparative analysis of tools in studies. A systematic literature review was conducted to evaluate the performance of nine popular and promising statistical and AI tools with a primary focus on 7 pertinent criteria (e.g., data size, error rate etc.) in the building energy research domain (see Chapter 3). one of the key contributions of this chapter is the development of a diagrammatic framework, carefully curated to serve as a guide for appropriate tool selection in various situations in the field of building energy consumption prediction.

In response to the initial framework founded through means of systematic literature review, the creation of a validated framework is an imperative step in the direction of enhancing the robustness and applicability of the tool for predicting energy consumption in diverse situations. The main purpose of this refined framework is to encompass the insights extrapolated from the empirical experiments highlighting the performance of tools when applied in various situations. Likewise, this strategic approach is tailored towards empowering Building Energy Prediction (BEP) model developers with a systematic guide for making well-informed decisions when choosing tools prediction, considering various criteria such as variable selection, and data size among others.

The empirical experiments conducted play a key role in validating the initial framework. Through these prior experiments, the performance of the different prediction tools is meticulously assessed in varying conditions. The presence of criteria like prediction type allows for a varying comprehension of how the predictive capabilities of various tools may differ across the different prediction types (i.e., classification, regression). Similarly, the type of variable selection becomes a crucial consideration, showcasing the relevance of specific variables in achieving accurate energy consumption prediction.

The number of variables or features is an often-overlooked phase in the academic literature, (Blum and Langley, 1997) which is systematically investigated in the experiments. This evaluation seeks to determine the optimal balance between inclusivity and simplicity of the predictive models, confirming that the chosen variables contribute implicitly to the performance of energy consumption predictions. Furthermore, the initial framework accounts for the influence of data size on model performance, acknowledging that model performance

may differ based on the availability of data. As a vital condition, the error rate was also considered to measure the predictive precision of the chosen tools. This measure helps provide a pragmatic comprehension of the reliability and trustworthiness of the prediction models which is fundamental for practical applications. Additionally, the evaluation of the computational cost of different tools is considered relevant knowledge, given that developers possess varying resource constraints. Balancing predictive accuracy with computational efficiency becomes paramount, especially in cases where large-scale deployment of BEP models is expected. Figure 8.1 shows the validated framework for energy consumption prediction of residential buildings.

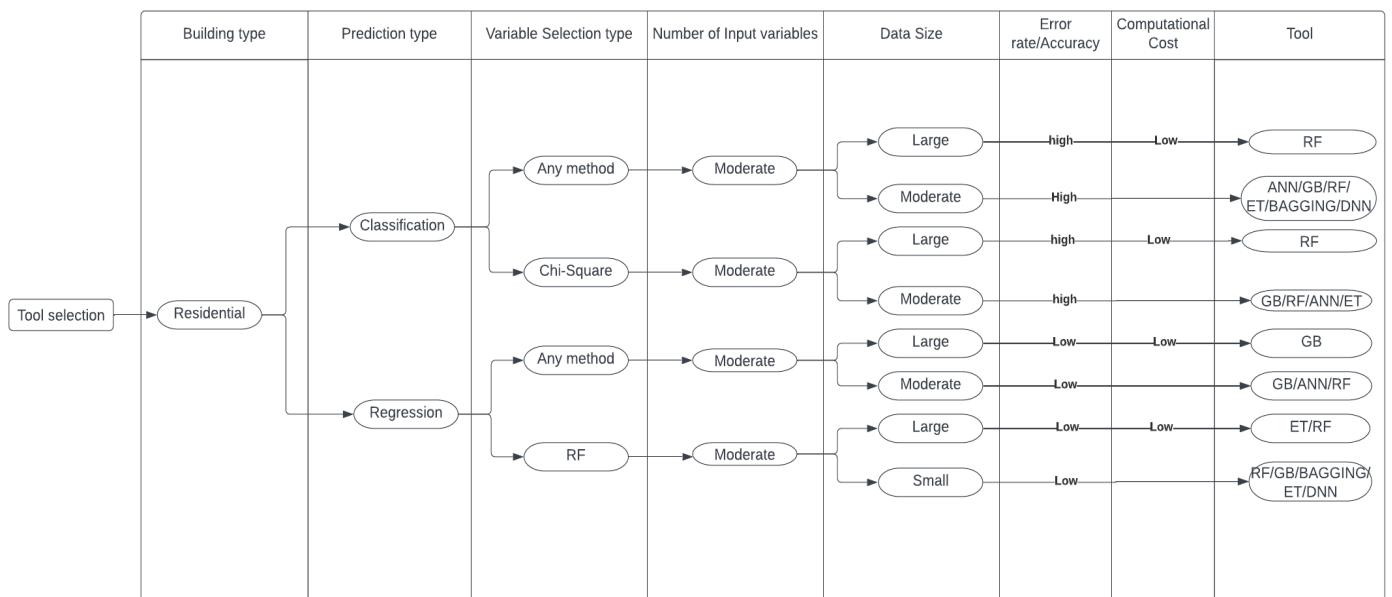


Figure 8.1: Validated framework

The systematic review of the literature offered results that were extrapolated for the creation of the initial framework for tool selection based on certain criteria such as data availability, error rate, and variable selection method(see section 3.8.8). It is suggested using Deep Neural Networks (DNN) for residential buildings with large data availability will produce low error rates, and Support Vector Machines (SVM) for cases with small data availability. However, this framework was further refined and validated through practical experimentation.

The validated framework, resulting from practical experimentation, presented additional deliberations and modifications to the initial recommendations or framework. For residential

buildings with large data availability, and a goal of achieving low error rates, and low computational constraints, Gradient Boosting (GB) was identified as the preferred choice over DNN due to its comparable performance with lower computational costs. This modification reveals a balance between accuracy and efficiency, acknowledging the resource constraints often combated in real-world applications. Therefore, the tool selection can differ depending on the developer's goal and resources. Moreover, while DNN engenders high computational costs which can pose practical challenges, it yields very good performance. It's not favourable in scenarios where computational resources are limited or cost-sensitive. Therefore, given that there is an alternative model that fulfils both criteria, the revised framework prioritizes models like GB that offer a favourable trade-off between computational efficiency and predictive performance.

Additionally, the scope of the framework extends its recommendations to cases with moderate data availability which is the case with real-world application (Ahmad et al., 2017b; C. Fan et al., 2017; Li et al., 2009a). In such contexts, the framework shows that multiple models like Artificial Neural Networks (ANN), Gradient Boosting (GB), and Random Forest (RF) produce good performance. By analysing multiple models, the framework proffers flexibility and robustness in tool selection for varying scenarios, accommodating data conditions and other requirements. Furthermore, this framework also includes the guidelines for tool selection in different prediction types (i.e., Classification, regression). Some model developers or researchers employ classification tasks for predicting energy efficiency ratings(Curtis et al., 2014; Iken et al., 2019), while some explore regression tasks for prediction of energy consumption values(Ahmed Gassar et al., 2019a; Diogo M. F. Izidio et al., 2021; Pham et al., 2020). This is to assess which produces prediction type elicits the best performance, as both will be beneficial for building designers and will revolutionize the development of more energy efficient buildings.

The framework explicitly depicts that the achievement of the best performance of a BEP model is determined by the appropriate selection of tools based on output type and characteristics of available data. This framework will ensure that tools are not selected based on popularity and unacademic factors. Also, BEP model developers will be able to select tools based on requirements. For instance, if a low computational cost is considered a top priority based on clients' specifications but the data available is large. DNN will not be wrongly selected rather, using this framework GB will be selected as the appropriate tool for such a situation.

Overall, the validated framework denotes a comprehensive and pragmatic method for tool selection for energy consumption prediction for residential buildings. By incorporating empirical findings and practical considerations, it provides actionable direction tailored to real-world scenarios, aiding informed decision-making and improving the effectiveness of energy consumption assessments in residential buildings.

The creation of a tool selection framework will empower statistical or machine learning model developers to make more informed decisions toward achieving model performance. Additionally, the machine learning model can empower designers to make well-informed decisions regarding energy-efficient design strategies, material selection, and system configurations in the early design phase. However, this research will go further by optimizing the algorithm by modifying the formula of the transparent statistical or machine learning algorithm.

8.3 Energy Optimizer

The development of an optimization model for energy consumption prediction encompasses the formulation of a well-defined objective function, which functions as the foundation of the predictive framework. This objective function captures the principal goal of the optimization process, intending to minimize prediction errors and enhance model performance or resource utilization(Gunantara, 2018). The choice of algorithm for developing an optimization model is predicated on the transparency of the model. Generally, ML models are considered black box models owing to their lack of transparency(Aggarwal et al., 2019), the most transparent tool is the statistical method namely Linear Regression, Logistic Regression among others. The regression model will be selected in this analysis because of its better performance than classification models as shown in Figure 7.23. Five statistical methods were used for model development however Linear Regression was selected based on its performance and prominence based on transparency in the literature(Biedma-Rdguez et al., 2022; Ciulla and D'Amico, 2019; Matveeva et al., 2007; Wood, 2022). Linear regression can create simple and transparent predictive models using equations arranged with their coefficients and bias in linear configuration (Wood, 2022).

Based on the transparency of the Linear Regression model, this LR objective function approach allows reverse engineering or back to front approach where a building designer can simply specify the desired or target energy value for a design, input it into the optimization model, and receive precise specifications for openings (i.e., windows, doors, etc.), configurations, total floor area, etc., essential to achieve the desired or target energy value. These algorithms operate in under five minutes and are poised to revolutionize the building construction process in a sustainable manner. The arduous and iterative nature of current design simulations to achieve marginal efficiency gains will soon become obsolete.

The implementation of this approach not only accelerates the design process but also significantly enhances the capacity of designers to create energy-efficient green buildings and environmentally friendly structures.

8.3.1 Implementation Strategy

To achieve this approach, Linear regression was selected. Although Linear regression (LR) did not produce the best performance in comparison to other model, it produces the best performance among the transparent algorithms. The selection of LR is owed to its interpretability, a critical factor in the context of optimization and formula modification.

The simplicity of LR allows for a clear and straightforward understanding of the relationships between input features and the target variable, making it an ideal choice for initial analysis and communication of results(Biedma-Rdguez et al., 2022; Wood, 2022). LR offers clear and intuitive insights into how each input feature impacts the predicted output. This transparency is highly desirable and beneficial in various fields, where stakeholders require a precise understanding of the factors that significantly drive the target output in order to make informed decisions. Unlike alternative regression models, which are often considered "black box" due to their complexity and lack of transparency, LR provides a clear mathematical representation that stakeholders can easily comprehend. However, despite the enhanced predictive performance offered by these more complex models, their opacity remains a significant drawback.

Furthermore, in the context of energy consumption prediction, by means of the LR formula, the coefficients (weights) and bias for each feature in the LR model can be utilized to determine the most optimal value for each feature. This also enables building designers to identify which building features have the most significant effect on energy consumption. This information empowers designers to focus attention on design modifications that proffer great potential for energy savings. The formula for linear regression can be used to further explain the interpretability of LR. The general formula for linear regression is (Borhani et al., 2022; Gopi, 2020):

$$Y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

Where:

Y is the predicted output,
 w_0 is the bias term (also known as the intercept),
 w_1, w_2, \dots, w_n are the coefficients (also known as weights) corresponding to each feature
 x_1, x_2, \dots, x_n

In this scenario, the user simply specifies the target or desired energy consumption value Y . The goal of the optimization process is to find the optimal values for the input features x_1, x_2, \dots, x_n such that, when inputted into the linear regression model, it will result in a predicted output Y is equal or close enough to the specified or desired target value Y .

However, this approach also offers the user the option to fix specific features, the optimization algorithm will adjust the feature values while keeping the feature fixed. This is done by minimizing the difference between the predicted output and the target value using an appropriate objective function. This fixed option is considered beneficial because not all building features are easily modifiable. For example, Figure 8.2 shows that the total floor area was fixed at 213 hence only optimal values for other features were generated. Furthermore, the weights and bias play different roles in this context, as explained below:

Weights (w_1, w_2, \dots, w_n): The weights denote the contribution of each feature to the predicted output (Xiao et al., 2017). In this scenario, the weight determines the magnitude in which each building feature influences the energy consumption prediction. During the optimization

process, the algorithm adjusts these weights to deduce the optimal values that produce the predicted output that is closest to the desired outcome or target value.

Bias (w_0): The bias term denotes the intercept of the linear regression model (Martin, 2000). It accounts for the baseline energy consumption when all building feature values are zero. In this context, the bias accounts for factors other than the input features that may have an effect on the energy consumption prediction. Like weights, the optimization algorithm modifies the bias term to ensure the predicted output aligns with the desired outcome or target value.

Generally, by leveraging the linear regression formula and optimization techniques, the user can specify a target energy consumption value, and the algorithm modifies the input features (excluding the building feature, which is fixed) to generate optimal feature values that elicit a predicted energy consumption value close to the target.

Therefore, in statistical term, given a target value \mathcal{Y} , the goal is to find the optimal values of all feature variables of x_n that satisfy the regression equation as closely as possible.

If the goal was to modify the formula to solve for one particular x_1 , while keeping the other x values fixed and solve for x_1 :

$$\mathcal{Y} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

$$\mathcal{Y} - w_0 + w_2x_2 + \cdots + w_nx_n = w_1x_1$$

$$x_1 = \frac{\mathcal{Y} - w_0 + w_2x_2 + \cdots + w_nx_n}{w_1}$$

However, in this case the optimal value for all features of x, this would typically require an optimization approach, especially if there's no single x value.

8.3.2 Optimization Approach

Optimization, in general, refers to the process of finding the best or optimal solution or outcome within a set of possible options, subject to defined constraints or goals (Pavlenko, 2019). Optimization is fundamental in various fields, from engineering and economics to computer

science, as it helps make decisions that yield the most efficient and effective results. In this application, optimization of the developed model focuses on minimizing energy consumption prediction errors by using a transparent Linear Regression model. This approach allows designers to specify a target energy value and receive precise building specifications to achieve that target.

To achieve the goal requires understanding of objective function. The objective function is the equation used to measure the difference between the predicted value and the target value (Zhang, 2021). The objective function $f(x)$ can be defined as:

$$f(x_1, x_2, \dots, x_n) = (Y - (w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n))$$

However, to minimize this difference (or error) between the predicted value and the target value requires the use of optimization algorithms. A commonly used method is the Nelder-Mead optimization (Ali and Tawhid, 2016; Chang, 2012; Ozaki et al., 2017). Nelder-Mead optimization is a search method used to minimize of an objective function of n-variable(Abdel-Basset et al., 2021).

This method starts with an initial estimate for the values of the feature variables x . This method then uses the optimization algorithm to adjust the values of x iteratively to minimize the objective function. The output of the optimization algorithm will be the set of feature values x_1, x_2, \dots, x_n that bring the predicted Y value as close as possible to the target Y as displayed Figure 8.2.

8.3.3 Energy Consumption Prediction Model Formula

The linear regression model formula for predicting the target variable (energy consumption), is as follows:

$$\begin{aligned} Y = & 740.4751 + (-58.4035 \cdot x_1) + (-0.3320 \cdot x_2) + (-27.5173 \cdot x_3) + (-27.5173 \cdot x_4) \\ & + (-16.9019 \cdot x_5) + (-16.9019 \cdot x_6) + (-73.6152 \cdot x_7) + (11.4060 \cdot x_8) \\ & + (25.5659 \cdot x_9) + (-10.8431 \cdot x_{10}) \end{aligned}$$

Where x_1, x_2, \dots, x_n are the features:

x_1 , MAINHEAT_DESCRIPTION | x_2 , TOTAL_FLOOR_AREA
 x_3 , WALLS_ENERGY_EFF | x_4 WALLS_ENV_EFF
 x_5 , ROOF_ENERGY_EFF | x_6 , ROOF_ENV_EFF
 x_7 , MAINHEAT_ENV_EFF | x_8 , FLOOR_DESCRIPTION
 x_9 , FLOOR_HEIGHT | x_{10} , AVG_TEMP

This formula combines various building and environmental feature, each highlighted by a specific coefficient that reflects its impact on the overall prediction. For instance, MAINHEAT_DESCRIPTION and MAINHEAT_ENV_EFF have significant negative coefficients, indicating that higher values in these features strongly reduce the target variable. Conversely, FLOOR_DESCRIPTION and FLOOR_HEIGHT have positive coefficients, suggesting they contribute to an increase in the target value. The intercept term or bias, 740.4751, represents the baseline value of the target variable when all features are zero. This comprehensive model encapsulates the complex relationship between different building features and their cumulative effect on energy consumption, providing a detailed analytical tool for predictive insights.

8.4 Reverse-Engineered System

This research developed a Reverse engineered system that allows for a back-to-front method for efficient energy assessment at the design stage of residential buildings, where a building designer can simply specify the desired or target energy value for a design, input it into the model, and receive precise specifications for each building feature, essential to achieve the desired or target energy value.

Essentially, the novelty of this approach lies in the ability to empower building designers with a methodical, data-driven mode to optimise energy use in building design. Currently, building designers rely on their simulation tools, domain knowledge, or expertise to make decisions around building features to ultimately achieve desired energy efficiency. However, the introduction of a more advanced strategy involving a combination of machine learning and statistical techniques with optimisation methods. By offering building designers an interface to input values for a set of building features, including the target or desired energy consumption and the values that needs to be fixed. This application offers a streamlined means of achieving design goals regardless of specific constraints.

The output of the deployed application includes the predicted energy consumption value and the optimal values for each building feature, representing a substantial contribution to building design. Therefore, not only does the application offer designers with potential energy consumption of their designs, but it also provides actionable recommendations for optimizing energy efficiency. Identification of the optimal values for each feature can be employed to achieve the target or desired energy consumption serve as a guide for designers towards more energy-efficient design solutions. Figure 8.2 shows a snapshot of the Reverse engineered system interface for energy assessment at the design stage.

Energy Optimizer

GreenOptimize - This is an energy prediction model that enables user to feature tune to achieve optimum energy performance

MAINHEAT_DESCRIPTION	4
TOTAL_FLOOR_AREA	213
WALLS_ENERGY_EFF	3
WALLS_ENV_EFF	3
ROOF_ENERGY_EFF	3
ROOF_ENV_EFF	3
MAINHEAT_ENV_EFF	2
FLOOR_DESCRIPTION	4
FLOOR_HEIGHT	68
AVG_TEMP	18
y_given	200
feature_to_fix_name	TOTAL_FLOOR_AREA

Clear **Submit**

output

('This is the predicted energy consumption value: [1611.33733076] and the optimal values to achieve target energy consumption are listed below:', Feature Optimal Value

0	MAINHEAT_DESCRIPTION	4.962558
1	WALLS_ENERGY_EFF	3.490109
2	WALLS_ENV_EFF	3.219619
3	ROOF_ENERGY_EFF	3.731719
4	ROOF_ENV_EFF	3.412661
5	MAINHEAT_ENV_EFF	2.243743
6	FLOOR_DESCRIPTION	3.975815
7	FLOOR_HEIGHT	12.105325
8	AVG_TEMP	15.499476)

Flag

Figure 8.2: Statistical and AI-based reverse engineering system

8.5 Chapter Summary

This chapter presents the validated framework and the reverse-engineered optimization model for energy for energy assessment at the design stage. The validated framework is then elaborated upon, delineating the variation between the validated and theoretical frameworks. Following this, the Energy Optimizer section discusses the implementation strategy, expatiating on the practical steps involved in the development and presents the formula of the linear regression model. This successful development of a reverse-engineered optimization model sets the groundwork for more effective energy assessment at the design stage, offering both theoretical and practical contributions.

CHAPTER NINE

9.0 CONCLUSIONS AND RECOMMENDATIONS

9.1 Chapter Overview

The chapter concludes this research and provides a comprehensive overview of the research. This chapter will review the research objectives and summarize key aspects such as main findings, contributions, limitations, and recommendations for future research. By reviewing these elements, this chapter captures the substance of the research journey and its implications on study and practice.

9.2 Review of Research Objectives and Conclusions

The primary aim of this research is to develop a reverse-engineered system for efficient energy assessment at the design stage of residential buildings. This allows for a back-to-front method where a building designer can simply specify the desired or target energy value for a design, input it into the model, and receive optimal value for each building feature, essential to achieve the desired or target energy value. To accomplish the stated aim, the following objectives were structured as follows.

1. To establish the key features that influence energy consumption in buildings using a systematic literature review.
2. To establish and validate the most prominent statistical and AI/ML algorithms for building energy consumption prediction in literature, using a systematic literature review.
3. To establish the minimum data size required for developing an efficient AI/ML energy prediction model, by benchmarking the performance of ML algorithms with varying data sizes.
4. To identify the most accurate and efficient building energy prediction model and the best performing transparent statistical and AI/ML models.
5. To develop an optimization model using the best performing transparent model from objective 4.

9.2.1 Objective One: To Establish the Key Features that Influence Energy Consumption in Buildings Using a Systematic Literature Review.

Q1: What are the most common features that influence energy consumption in buildings?

The achievement of effective outcomes from the several methods being employed to improve building energy efficiency requires an understanding of the factors influencing energy consumed in buildings. Numerous BEPMs have been developed to improve building energy efficiency however majority have developed using factors solely based on their popularity without a proper understanding of the effect and impact on energy consumption. Unfortunately, these have engendered unstable models as they omit some essential BEP factors.

This research implemented a systematic literature review method to highlight the most pertinent factors influencing BEP. From the evidence in Chapter 4, it can be concluded that three key factors namely weather, walls and windows are the most important factors. Though not popularly researched in the reviewed articles, the roof appeared to be a very essential factor affecting energy use in buildings and it should be adequately considered in the development of an efficient BEPM. The research evidently showed that the identified factors cut across the different types of buildings and climatic conditions around the world which makes them more relevant to developing reliable and holistic or generalisable BEPM. Chapter 4 fulfils this objective for this research, highlighting the impact of each identified factors important factors impact on BEP. Additionally, it discussed several recommendations of energy efficient strategies for building designers such as the utilization of green roofs, and photovoltaic (PV) windows among others which have been considered to have significant effects on BEP.

9.2.2 Objective Two: To Establish and Validate the Most Prominent Statistical and AI/ML Algorithms for Building Energy Consumption Prediction Using a Systematic Literature Review.

Q2: What are the most prominent statistical or AI/ML algorithms for energy consumption prediction?

Research in the BEP field continues to increase with the development of several new models using diverse tools. Nevertheless, the majority of these tools are employed in unsuitable data conditions or for the wrong situation. Based on different pertinent criteria, this research conducted a systematic literature review of common and promising tools in the BEP domain. Overall, from the result in Chapter 3, it is concluded that no singular tool is primarily better

than all other tools across the identified criteria and emphasizes that no one tool is suitable for all purposes under different circumstances. Thus, it is evident that All tools have their strengths and drawbacks and produce different outcomes under different circumstances (i.e., data conditions, and developer goals, among others). This research provided a simplified framework that will avail developers the opportunity to make more informed decisions when selecting tool(s) most suitable for the situation or condition, rather than make selections centred on popularity and unacademic factors, thus, tools will be selected often based on their respective strengths. This research will alleviate the exhaustive process of developing several BEP models using various tools at random or based on popularity to select the most suitable for the developer's purpose.

The research revealed that certain research areas may need additional attention: annual and monthly energy use prediction, energy consumption prediction for residential buildings and natural gas energy prediction. The relatively low research focus in these areas could be subject to the data inadequacies and/or intricacies of occupant energy use behaviour. The availability of sufficient data (in terms of types, sizes and temporal granularities) is critical for enhanced energy consumption prediction and increased research efforts in these areas. Various buildings have been equipped with smart meters which will lead to large data sizes, the provision of these data on repositories available to researchers will engender unparalleled research in understanding energy efficiency in buildings, which will also eradicate the repetition of effort without unprecedented outcomes. Further to the validation, it was noted that, although GB has not received much attention in the field of energy performance prediction, the performance level of the ML algorithm proffers GB as an effective predictive model in the field of energy prediction.

9.2.3 Objective Three: To Establish the Minimum Data Size Required for Developing an Efficient AI/ML Energy Prediction Model, by Benchmarking the Performance of ML Algorithms with Varying Data Sizes

Q3: What is the minimum data size required for efficient energy prediction at the design stage of buildings?

To establish the minimum data size required for energy prediction at the design stage of buildings, this research conducted a reliability analysis. It can be inferred from the outcome of the analysis, that while models developed using smaller datasets (20% and 40% of the data)

exhibit acceptable performance, the optimal performance was achieved with GB on a larger dataset (100% data availability). Generally, it is concluded that the performance of the models improves as the dataset size increases, denoting that more data leads to more accurate energy predictions. However, it's important to note that even with smaller datasets, some models produced relatively good performance for example, RF achieved r^2 of 0.79 at 20% data availability, albeit not reaching the optimal performance observed with the full dataset. Therefore, while there isn't a specific minimum data size identified, it can be established that employing as much data as possible, is necessary for developing an optimal energy prediction model at the design stage of buildings.

9.2.4 Objective Four: To Identify the Most Accurate and Efficient Building Energy Prediction Model and the Best Performing Transparent Statistical and AI/ML Models

Q4: Which statistical or AI/ML algorithm is the most accurate, efficient, and transparent for predicting potential energy consumption at the design stage of buildings?

Based on the comparative analysis conducted, the Gradient Boosting (GB) model emerged as the most effective algorithm for predicting potential energy consumption at the design stage of buildings, achieving a striking R-squared value of 0.86. Despite its longer runtime of 44.2 seconds, the high r^2 value of the GB model makes it a compelling option for scenarios where precision is paramount. However, in terms of efficiency, the Lasso model outperformed all other models, producing a decent R-squared value of 0.70 in 0.11 seconds. This highlights Lasso's ability to produce results swiftly, making it a compelling option for cases where computational resources are limited. Although the Deep Neural Network (DNN) model displayed competitive performance with an R-squared of 0.83, its significantly longer runtime of 5715.613 seconds makes it less practical in comparison to GB. Therefore, despite its relatively long runtime and considering it is significantly faster than the traditional method of energy assessment at the design stage, the Gradient Boosting model is concluded as the optimal choice, striking a balance between accuracy and efficiency for predicting building energy consumption at the design stage.

9.2.5 Objective Five: To Develop an Optimization Model Using the Best Performing Transparent Model from Objective 4.

Q5: What are the key components and parameters necessary to develop an optimization model?

The development of an optimization model was conducted using linear regression which was identified as one of the most transparent from literature. In this case, where the target is to input a dependent variable and return optimal values for the independent variables, it is concluded that the key components and parameters include the dependent variable (target outcome to be achieved), independent variables (feature whose optimal values are required), and the optimization algorithm. This is discussed more elaborately in section 8.3.2.

By structuring these objectives effectively, this led to successfully achievement of the primary aim of this research. This research developed a reverse-engineered energy assessment system for residential building design. This system allows designers to specify a target energy value and receive optimal values for each relevant building feature, enabling them to meet the desired energy outcome.

9.3 Contributions and Significance of the Research

This section presents the theoretical and practical implications of this research. This research elucidates significant theoretical and practical implications, enhancing both academic understanding and real-world application. Theoretically, it contributes to the existing body of knowledge by introducing novel insights and frameworks that refine established paradigms, offering a deeper comprehension of the subject matter. Practically, the findings provide actionable strategies and solutions that can be implemented in practice, improving efficiency, decision-making processes, and overall outcomes. By bridging the gap between theory and practice, this research not only advances scholarly discourse but also equips practitioners with evidence-based tools to address contemporary challenges effectively.

9.3.1 Theoretical Implications of Research

This research delivers several theoretical contributions. One of the significant contributions of this research is the provision of a simplified framework that enables developers to make more informed decisions when selecting tools, focusing on suitability for specific situations rather

than on popularity or other unacademic factors. This approach ensures that tools are chosen based on their respective strengths and applicability, leading to more effective and efficient outcomes. Over the past few years, the selection of tools for building energy prediction has typically been done arbitrarily or based solely on their popularity, without taking into account their strengths and weaknesses in certain conditions (e.g., Divina et al., 2018b; Feng and Zhang, 2020; D.M.F. Izidio et al., 2021; C. Robinson et al., 2017). This approach has engendered poor performance and time-consuming comparative analysis of tools in studies. This research conducted a systematic literature review of popular and favourable building energy consumption prediction tools and based on the findings, created a framework to facilitate a well-informed selection of tools. However, notably results from theory could differ from practical based on minor discrepancies in data. Therefore, this research further conducted empirical analysis to validate the findings from the review of the literature and generated a validated framework. This will essentially lead to well-informed utilization of tools for building energy consumption prediction. This research will engender a fundamental shift from the exhaustive process of developing several BEP models using various tools selected at random or based on popularity to selecting the most suitable tool for the developer's purpose. Moreover, by analysing the performance of statistical and AI-based tools across different conditions and identifying their strengths and limitations, this research contributes to the development of a better understanding of the effectiveness of these tools in different scenarios. Different developers attempt to develop models with a goal in mind, however, the achievement of this goal is predicated on the conditions. For example, accuracy is of great importance when developing a BEP model however the accuracy rate is highly correlated to the input/output, and data size amongst other conditions(Fathi et al., 2020b; Goyal et al., 2020; Runge and Zmeureanu, 2019). The research provides insights into the factors that influence the performance of these tools, which can inform the development of new hybrid models and improve the accuracy of building energy prediction models.

Nonetheless, this research offers many theoretical contributions by addressing several contemporary hypotheses in academic literature (see section 1.4.2). For example, feature selection is proclaimed to have a significant effect on model performance and some studies have argued that feature selection(FS) is more effective in classification than regression prediction (Jović et al., 2015; Kumar, 2014), this research explored the utilization of various FS techniques in developing several statistical and machine learning models for different prediction type(Classification, regression) to conduct a clear and unbiased comparison. Some

studies suggest that feature selection is essential for the optimum performance of the model (Alaka et al., 2018; Balogun et al., 2021; Zhang and Wen, 2019a; Hai-xiang Zhao and Magoulès, 2012a); some studies suggest that feature selection is unfavourable to certain algorithms (Alaka et al., 2019, 2018; Olu-Ajayi et al., 2022a). However, this research showed that feature selection can have a positive or negative impact on predictive accuracy, depending on the algorithm selected. Thus, the achievement of a good predictive accuracy of a model is highly predicated on the type of algorithm and feature selection method chosen (Olu-Ajayi et al., 2022b). Therefore, if an appropriate feature selection method is not selected for the specific ML algorithm used, feature selection can result in poor performance. Overall, the choice of the most suitable feature selection method can help in identifying the most relevant features that contribute to the energy use of a building. This research identified the most suitable FS method for specific statical and ML algorithms. In terms of technical implications, technical expertise is required to implement different feature selection methods and explore various FS methods to identify the most suitable for an ML algorithm which is more labour-intensive. This research delivers the most suitable FS method for certain ML algorithms, reducing the time-consuming process of exploring various FS methods.

In literature, another hypothesis is the idea that weather data improves performance for energy prediction without a clear and unbiased comparison of the model performance with and without weather data to validate the effect of weather data on model performance. This research noted that, although the addition of weather data does improve the performance, it does not engender a significant increase.

One general hypothesis in the machine learning world, states that the larger the data utilized for model training, the better the performance (Dalal, 2018; Goyal et al., 2020; Kaur and Gupta, 2017; Lee et al., 2011). In this research, a reliability analysis was conducted to substantiate or disprove this hypothesis. Different machine learning models were developed using five sample sizes (20%, 40%, 60%, 80%, 100%). This finding engendered from the analysis suggests that the larger the data does not always lead to a better result. Although, notably, a single study is not enough to substantiate this conclusion, and this should be subject to further investigation. Some studies have corroborated certain conclusions, for example, Deep Neural Networks (DNN) showed a relative increase in predictive accuracy based on the increase in data size. Past studies have shown that large data has an effect on certain machine learning algorithms such as neural networks, (Amasyali and El-Gohary, 2018; Bourhnane et al., 2020). Also, the

study by Bourhnane et al., 2020 stipulates that neural network algorithms are dominant with big datasets, as they require sufficient data to train the model. On the other hand, SVC shows no significant increase in predictive accuracy based on the increase in the data size. This is can be subject to the conclusion that Support Vector Machines (SVM) are recognized for their ability to deliver good results effectively regardless of data size (Li et al., 2009; Qiong Li et al., 2010). Therefore, the general hypothesis on large data and better predictive performance is only peculiar to certain machine learning algorithms.

Finally, another major theoretical contribution is the development of Linear Regression optimization models which generate the optimal values of each feature that contribute to achieving a target energy consumption level. This not only helps in comprehension of the relative importance of different features but also leads to the development of more targeted strategies for energy efficiency in various domains, ranging from building management to industrial operations, thereby contributing to sustainability efforts and cost reduction initiatives.

9.3.2 Practical Implications of Research

This research has significant practical implications, which are poised to revolutionize the building construction process in a sustainable manner by offering building designers an interface to input values for a set of building features, including the target or desired energy consumption and the value that needs to be fixed. The reverse-engineered system developed offers a back-to-front approach where a building designer will receive optimal feature values for each feature required to achieve the precise of close enough to the target energy consumption. This system operates in under five minutes, a significant improvement over the typical timeframe, which can range from several hours to weeks depending on various factors. This significant reduction in processing time has the potential to greatly enhance productivity.

At the design stage, the development of a prediction model with excellent performance would provide great support for building designers. It will enable designers to deduce the potential energy use of a building at the design stage, optimize design instantaneously based on the energy predicted and conduct continuous iteration until optimum performance is achieved. This will potentially reduce the construction of energy-inefficient buildings. This research developed an AI/ML model for predictions of the potential energy consumption at the design

stage. This model will be used by a building designer, to simply input values for features related to the design such as total floor area (i.e., 20m²). Once these values have been inputted, the designer will run the model and receive the result in less than five minutes. The designer will receive the potential energy consumption at the design. If the energy value predicted doesn't meet the requirements, the building designer will modify the key variables and run the model again iteratively, until optimal energy performance is achieved.

Additionally, this research develops a reverse-engineered system to further streamline the iteration process and enhance efficiency. This research developed a reverse engineered system that enables building designers to specify the target energy consumption value and receive the optimal values for each feature, thus limiting the continuous iteration of features to achieve target energy consumption. This would potentially decrease the construction of more energy-efficient buildings that are detrimental to the environment.

One of the key advantages of using an AI model at the design stage is the ability to perform faster iterative processes. Nonetheless, this research goes a step further by reducing the iteration time even more. The novel aspect of this research is the introduction of a reverse-engineered system that allows designers to input values for each feature and the desired or target energy consumption value into the system. Once the AI model predicts the potential energy consumption based on the design feature values supplied, it returns the optimal values for each design feature to achieve the target energy consumption value. This approach minimizes the need for repeated iterations, as the system directly provides the adjustments needed to meet the target energy consumption. This not only saves time but also enhances the precision and effectiveness of the design process. This system will certainly help expedite the work of designers

Currently, building designers rely on their domain knowledge, intuition or expertise to make decision around building features to ultimately achieve desired energy efficiency. However, the introduction of a more advanced strategy involving a combination of statistical and machine learning techniques with optimisation algorithms that proffers the optimal value will be tremendously beneficial to building designers and other stakeholders. This method will aid a significant advancement in the development of more energy-efficient buildings by bridging the gap between design intent and energy performance outcomes. By employing the reverse-engineered system in the design process, the application allows designers to make informed

decisions that prioritize energy efficiency without typically compromising on the design quality or functionality. Ultimately, this contributes to the wider goal of sustainable architecture or green buildings by supporting the development of buildings that ameliorate environmental impact while amplifying occupant comfort and well-being.

Finally, as we aspire to create a more sustainable future and endeavour to diminish carbon footprint, optimizing energy usage in buildings assumes increasing importance. Although conducting an energy evaluation during the design phase entails a substantial investment of time and resources which can be quite frustrating and discouraging for designers, and the reliance on previous experience often leads to inadequate conclusions. This developed reverse-engineered system will open doors to remarkable productivity, promote informed decision making and eliminate frustration for designers. Additionally, property marketers can benefit from predicted building energy consumption levels as it constitutes a means to increase selling costs to customers searching for energy-efficient properties. Consequently, this could encourage property developers and building consumers to consider energy performance before development or acquisition. From an environmental standpoint, this is expected to have a predominantly positive impact, as it promotes energy-efficient building designs contributes to a reduction in greenhouse gas emissions and minimizes resource consumption.

9.4 Limitations of the Research

The research has a few limitations that should be acknowledged. Firstly, this research focuses exclusively on residential building data, which restricts the generalizability of the findings, making them less applicable to commercial or other types of buildings. Notably, energy consumption studies often emphasize commercial buildings, as shown in Figure 4.8, despite commercial buildings representing a small proportion of all buildings. This is because, unlike residential buildings, commercial buildings typically have Building Energy Management Systems (BEMS) that facilitate data collection which avails data for analysis. Additionally, this research is limited to annual energy consumption, which does not capture the different temporal granularities of energy usage within buildings. This focus is due to the research's emphasis on energy prediction at the design stage, using building feature values to predict potential energy consumption, unlike studies on existing or operational buildings that use temporal granular data such as hourly or weekly data for future predictions. Lastly, energy performance data while useful for developing an energy prediction model could potentially have some notable

limitations that could impact their reliability considering the assessment for energy performance certificate requires human intervention. To address these limitations, future research should explore the use of IoT devices to capture more accurate energy performance data.

9.5 Recommendations for Future Research

Future research should explore several other criteria and conduct practical experiments to further substantiate or nullify findings for other granularities such as hourly, and daily energy prediction. Future research should explore other contemporary Deep learning algorithms which have not been employed in literature, given that Deep Neural Networks (DNNs) produced good outcomes and met various performance criteria, particularly across different data sizes. While models like Long Short-Term Memory (LSTM) and Autoregressive Integrated Moving Average (ARIMA) have been extensively studied in the literature for time series prediction, they appear less suitable for annual energy prediction. Thus, further exploration of innovative deep-learning approaches could yield significant advancements in the accuracy and applicability of energy forecasting models.

Although it is inferred that no singular tool performs best in all criteria, future research should explore the integration of multiple statistical and AI-based prediction tools to improve the accuracy of building energy use prediction. The identification of the strengths and limitations of different tools may inform the development of hybrid models that leverage the strengths of each approach, this could engender hybrid tools that can achieve good performance in all or most of the identified criteria.

Linear Regression (LR) was selected for optimization in this research primarily due to its interpretability and transparency. The interpretability of LR emanates from its simplicity and transparency in demonstrating the relationship between input features and the target variable. Although alternative regression models are often considered "black box" due to their complexity and lack of transparency, the enhanced predictive performance offered by these more complex models warrants further investigation. Therefore, future research should focus on developing methods to demystify these black-box models, aiming to combine their superior

predictive capabilities with improved interpretability, ultimately leading to more robust and comprehensible analytical tools.

Additionally, modification of the objective functions to incorporate domain-specific constraints or preferences could potentially tailor the optimization process to better suit specific contexts, ensuring alignment with real-world objectives and constraints. Furthermore, integrating advanced optimization techniques such as metaheuristic algorithms or evolution optimization algorithms could offer more sophisticated and flexible solutions for tackling complex optimization problems related to energy consumption. Overall, by diversifying modelling approaches and refining optimization strategies, future research can advance the comprehension and capabilities in optimizing energy consumption, leading to more sustainable and efficient energy usage practices across various domains.

Future research could benefit from encompassing a wider range of building types and exploring more granular energy consumption data to enhance the comprehensiveness.

9.6 Chapter Summary

This chapter draws together the findings from the research and delivers conclusions and recommendations for future research. This chapter addresses each objective individually, to demonstrate the fulfilment of all objectives. Subsequently, the chapter highlights the contributions and significance of the research, both theoretically and practically. This research developed a back-to-front model which enables building designers to assess energy more effectively at the design stage, which is considered one of the most significant contributions in this research. Furthermore, this chapter acknowledges the limitations of this research and offers recommendations for future research, suggesting areas for further investigation and refinement.

REFERENCES

- Aadithyan, V., Goud, T.S.S., Reddy, G.K., Chaitanya, P.N., Surya, V.J., Rao, K.P., 2020. Smart Face Recognition System.
- Abdel-Basset, M., Mohamed, R., Mirjalili, S., 2021. A novel Whale Optimization Algorithm integrated with Nelder–Mead simplex for multi-objective optimization problems. *Knowledge-Based Systems* 212, 106619. <https://doi.org/10.1016/j.knosys.2020.106619>
- Abdollahi, H., Mofid, B., Shiri, I., Razzaghdoost, A., Saadipoor, A., Mahdavi, A., Galandooz, H.M., Mahdavi, S.R., 2019. Machine learning-based radiomic models to predict intensity-modulated radiation therapy response, Gleason score and stage in prostate cancer. *Radiol med* 124, 555–567. <https://doi.org/10.1007/s11547-018-0966-4>
- Aboelata, A., 2021. Assessment of green roof benefits on buildings' energy-saving by cooling outdoor spaces in different urban densities in arid cities. *Energy* 219. <https://doi.org/10.1016/j.energy.2020.119514>
- Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., Calhoun, V., 2021. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat Commun* 12, 353. <https://doi.org/10.1038/s41467-020-20655-6>
- Adegoke, M., Hafiz, A., Ajayi, S., Olu-Ajayi, R., 2022. Application of Multilayer Extreme Learning Machine for Efficient Building Energy Prediction. *Energies* 15, 9512. <https://doi.org/10.3390/en15249512>
- Aggarwal, A., Lohia, P., Nagar, S., Dey, K., Saha, D., 2019. Black box fairness testing of machine learning models, in: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2019. Association for Computing Machinery, New York, NY, USA, pp. 625–635. <https://doi.org/10.1145/3338906.3338937>
- 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>

- Ahmad, M.W., Mouraud, A., Rezgui, Y., Mourshed, M., 2018. Deep highway networks and tree-based ensemble for predicting short-term building energy consumption. *Energies* 11. <https://doi.org/10.3390/en11123408>
- Ahmad, M.W., Mourshed, M., Rezgui, Y., 2017a. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings* 147, 77–89.
<https://doi.org/10.1016/j.enbuild.2017.04.038>
- Ahmad, M.W., Mourshed, M., Rezgui, Y., 2017b. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings* 147, 77–89.
<https://doi.org/10.1016/j.enbuild.2017.04.038>
- Ahmed Gassar, A.A., Yun, G.Y., Kim, S., 2019a. Data-driven approach to prediction of residential energy consumption at urban scales in London. *Energy* 187, 115973.
<https://doi.org/10.1016/j.energy.2019.115973>
- Ahmed Gassar, A.A., Yun, G.Y., Kim, S., 2019b. Data-driven approach to prediction of residential energy consumption at urban scales in London. *Energy* 187, 115973.
<https://doi.org/10.1016/j.energy.2019.115973>
- Ahn, K.U., Kim, D.W., Kim, Y.J., Yoon, S.W., Park, C.S., 2016. Issues to be solved for energy simulation of an existing office building. *Sustainability (Switzerland)* 8.
<https://doi.org/10.3390/su8040345>
- Akbar, B., Amber, K.P., Kousar, A., Aslam, M.W., Bashir, M.A., Khan, M.S., 2020. Data-driven predictive models for daily electricity consumption of academic buildings. *AIMS Energy* 8, 783–801. <https://doi.org/10.3934/ENERGY.2020.5.783>
- Alaka, H., Oyedele, L., Owolabi, H., Akinade, O., Bilal, M., Ajayi, S., 2019. A Big Data Analytics Approach for Construction Firms Failure Prediction Models. *IEEE Trans. Eng. Manage.* 66, 689–698. <https://doi.org/10.1109/TEM.2018.2856376>
- Alaka, H.A., Oyedele, L.O., Owolabi, H.A., Kumar, V., Ajayi, S.O., Akinade, O.O., Bilal, M., 2018. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications* 94, 164–184.
<https://doi.org/10.1016/j.eswa.2017.10.040>
- Alawadi, S., Mera, D., Fernández-Delgado, M., Alkhabbas, F., Olsson, C.M., Davidsson, P., 2020. A comparison of machine learning algorithms for forecasting indoor temperature in smart buildings. *Energy Systems*. <https://doi.org/10.1007/s12667-020-00376-x>

- Alduailij, Mona A., Petri, I., Rana, O., Alduailij, Mai A., Aldawood, A.S., 2021. Forecasting peak energy demand for smart buildings. *J Supercomput* 77, 6356–6380.
<https://doi.org/10.1007/s11227-020-03540-3>
- Al-Homoud, M.S., 2001. Computer-aided building energy analysis techniques. *Building and Environment* 36, 421–433. [https://doi.org/10.1016/S0360-1323\(00\)00026-3](https://doi.org/10.1016/S0360-1323(00)00026-3)
- Ali, A.F., Tawhid, M.A., 2016. A hybrid cuckoo search algorithm with Nelder Mead method for solving global optimization problems. *SpringerPlus* 5, 473.
<https://doi.org/10.1186/s40064-016-2064-1>
- Ali, N., Neagu, D., Trundle, P., 2019. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Appl. Sci.* 1, 1559.
<https://doi.org/10.1007/s42452-019-1356-9>
- Allouhi, A., El Fouih, Y., Kousksou, T., Jamil, A., Zeraouli, Y., Mourad, Y., 2015. Energy consumption and efficiency in buildings: current status and future trends. *Journal of Cleaner Production, Special Issue: Toward a Regenerative Sustainability Paradigm for the Built Environment: from vision to reality* 109, 118–130.
<https://doi.org/10.1016/j.jclepro.2015.05.139>
- Almalaq, A., Zhang, J.J., 2019. Evolutionary Deep Learning-Based Energy Consumption Prediction for Buildings. *IEEE Access* 7, 1520–1531.
<https://doi.org/10.1109/ACCESS.2018.2887023>
- Al-Rakhami, M., Gumaei, A., Alsanad, A., Alamri, A., Hassan, M.M., 2019. An Ensemble Learning Approach for Accurate Energy Load Prediction in Residential Buildings. *IEEE Access* 7, 48328–48338. <https://doi.org/10.1109/ACCESS.2019.2909470>
- Al-Shargabi, B., Al-Shami, F., 2019. An experimental study for breast cancer prediction algorithms, in: *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems, DATA '19*. Association for Computing Machinery, New York, NY, USA, pp. 1–6. <https://doi.org/10.1145/3368691.3368703>
- Alwetaishi, M., Benjeddou, O., 2021. Impact of window to wall ratio on energy loads in hot regions: A study of building energy performance. *Energies* 14.
<https://doi.org/10.3390/en14041080>
- Alwetaishi, Mamdooh, Benjeddou, O., 2021. Impact of Window to Wall Ratio on Energy Loads in Hot Regions: A Study of Building Energy Performance. *Energies* 14, 1080.
<https://doi.org/10.3390/en14041080>

- Amasyali, K., El-Gohary, N., 2021. Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. *Renewable and Sustainable Energy Reviews* 142. <https://doi.org/10.1016/j.rser.2021.110714>
- Amasyali, Kadir, El-Gohary, N., 2021. Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. *Renewable and Sustainable Energy Reviews* 142, 110714. <https://doi.org/10.1016/j.rser.2021.110714>
- Amasyali, K., El-Gohary, N., 2017. Deep Learning for Building Energy Consumption Prediction.
- Amasyali, K., El-Gohary, N.M., 2018. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews* 81, 1192–1205. <https://doi.org/10.1016/j.rser.2017.04.095>
- Amber, K.P., Ahmad, R., Aslam, M.W., Kousar, A., Usman, M., Khan, M.S., 2018. Intelligent techniques for forecasting electricity consumption of buildings. *Energy* 157, 886–893. <https://doi.org/10.1016/j.energy.2018.05.155>
- Amiri, F., Rezaei Yousefi, M., Lucas, C., Shakery, A., Yazdani, N., 2011. Mutual information-based feature selection for intrusion detection systems. *Journal of Network and Computer Applications, Advanced Topics in Cloud Computing* 34, 1184–1199. <https://doi.org/10.1016/j.jnca.2011.01.002>
- Amsterdamska, O., Leydesdorff, L., 2005. Citations: Indicators of significance? *Scientometrics* 15, 449–471. <https://doi.org/10.1007/bf02017065>
- Andargie, M.S., Touchie, M., O'Brien, W., 2019. A review of factors affecting occupant comfort in multi-unit residential buildings. *Building and Environment* 160, 106182. <https://doi.org/10.1016/j.buildenv.2019.106182>
- Ang, J.C., Mirzal, A., Haron, H., Hamed, H.N.A., 2016. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13, 971–989. <https://doi.org/10.1109/TCBB.2015.2478454>
- AR5 Synthesis Report: Climate Change — IPCC [WWW Document], 2014. URL <https://www.ipcc.ch/report/ar5/syr/> (accessed 8.10.22).
- Ardjmand, E., Millie, D.F., Ghalehkondabi, I., Young II, W.A., Weckman, G.R., 2016. A State-Based Sensitivity Analysis for Distinguishing the Global Importance of Predictor Variables in Artificial Neural Networks. *Advances in Artificial Neural Systems* 2016, e2303181. <https://doi.org/10.1155/2016/2303181>

- Asilevi, P.J., Quansah, E., Amekudzi, L.K., Annor, T., Klutse, N.A.B., 2019. Modeling the spatial distribution of Global Solar Radiation (GSR) over Ghana using the Ångström-Prescott sunshine duration model. *Scientific African* 4, e00094.
<https://doi.org/10.1016/j.sciaf.2019.e00094>
- Asir, D., Gnana, A., Leavline, E.J., 2016. Literature Review on Feature Selection Methods for High-Dimensional Data.
- Athey, S., 2019. 21. The Impact of Machine Learning on Economics, in: 21. The Impact of Machine Learning on Economics. University of Chicago Press, pp. 507–552.
<https://doi.org/10.7208/9780226613475-023>
- Aversa, P., Donatelli, A., Piccoli, G., Luprano, V.A.M., 2016. Improved Thermal Transmittance Measurement with HFM Technique on Building Envelopes in the Mediterranean Area. *Selected Scientific Papers - Journal of Civil Engineering* 11.
<https://doi.org/10.1515/sspjce-2016-0017>
- Aziz, R., Verma, C.K., Srivastava, N., Aziz, R., Verma, C.K., Srivastava, N., 2017. Dimension reduction methods for microarray data: a review. *AIMSBOA* 4, 179–197.
<https://doi.org/10.3934/bioeng.2017.1.179>
- Badiei, A., Akhlaghi, Y.G., Zhao, X., Li, J., Yi, F., Wang, Z., 2020. Can whole building energy models outperform numerical models, when forecasting performance of indirect evaporative cooling systems? *Energy Conversion and Management* 213.
<https://doi.org/10.1016/j.enconman.2020.112886>
- Bagnasco, A., Fresi, F., Saviozzi, M., Silvestro, F., Vinci, A., 2015. Electrical consumption forecasting in hospital facilities: An application case. *Energy and Buildings* 103, 261–270. <https://doi.org/10.1016/j.enbuild.2015.05.056>
- Bahassine, S., Madani, A., Al-Sarem, M., Kissi, M., 2020. Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences* 32, 225–231.
<https://doi.org/10.1016/j.jksuci.2018.05.010>
- Balogun, H., Alaka, H., Egwim, C.N., 2021. Boruta-grid-search least square support vector machine for NO₂ pollution prediction using big data analytics and IoT emission sensors. *Applied Computing and Informatics ahead-of-print*.
<https://doi.org/10.1108/ACI-04-2021-0092>
- Barakat, E.H., Qayyum, M.A., Hamed, M.N., Rashed, S.A., 1990. Short-term peak demand forecasting in fast developing utility with inherit dynamic load characteristics. I. Application of classical time-series methods. II. Improved modelling of system

- dynamic load characteristics. Power Systems, IEEE Transactions on 5, 813–824. <https://doi.org/10.1109/59.65910>
- Barboza, F., Kimura, H., Altman, E., 2017. Machine learning models and bankruptcy prediction. Expert Systems with Applications 83, 405–417. <https://doi.org/10.1016/j.eswa.2017.04.006>
- Barton, C., Wilson, W., 2021. Tackling the under-supply of housing in England.
- Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks 5, 537–550. <https://doi.org/10.1109/72.298224>
- Bayona-Oré, S., Cerna, R., Hinojoza, E.T., 2021. Machine Learning for Price Prediction for Agricultural Products. WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS 18, 969–977. <https://doi.org/10.37394/23207.2021.18.92>
- BEIS [WWW Document], 2019. . GOV.UK. URL <https://www.gov.uk/government/publications/beis-annual-report-and-accounts-2019-to-2020> (accessed 9.15.21).
- Benhar, H., Idri, A., Fernández-Alemán, J.L., 2020. Data preprocessing for heart disease classification: A systematic literature review. Computer Methods and Programs in Biomedicine 195, 105635. <https://doi.org/10.1016/j.cmpb.2020.105635>
- Berardi, U., Tronchin, L., Manfren, M., Nastasi, B., 2018. On the effects of variation of thermal conductivity in buildings in the Italian construction sector. Energies 11. <https://doi.org/10.3390/en11040872>
- Berk, R.A., 2006. An Introduction to Ensemble Methods for Data Analysis. Sociological Methods & Research 34, 263–295. <https://doi.org/10.1177/0049124105283119>
- Beykzade, M., Beykzade, S., 2019. EXAMINING THE COMPONENTS OF GREEN BUILDING DESIGN AND ITS MANAGEMENT SYSTEM. Eurasian Journal of Civil Engineering and Architecture 3, 48–52.
- Bhattacharjee, S., Reichard, G., 2011. Socio-Economic Factors Affecting Individual Household Energy Consumption: A Systematic Review. Presented at the ASME 2011 5th International Conference on Energy Sustainability, ES 2011. <https://doi.org/10.1115/ES2011-54615>
- Biedma-Rdguez, C., Gacto, M.J., Anguita-Ruiz, A., Alcalá-Fdez, J., Alcalá, R., 2022. Transparent but Accurate Evolutionary Regression Combining New Linguistic Fuzzy Grammar and a Novel Interpretable Linear Extension. Int. J. Fuzzy Syst. 24, 3082–3103. <https://doi.org/10.1007/s40815-022-01324-w>

Bijarniya, J.P., Sarkar, J., Maiti, P., 2020. Environmental effect on the performance of passive daytime photonic radiative cooling and building energy-saving potential. Journal of Cleaner Production 274. <https://doi.org/10.1016/j.jclepro.2020.123119>

Blagus, R., Lusa, L., 2015. Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics 16, 363. <https://doi.org/10.1186/s12859-015-0784-9>

Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106. <https://doi.org/10.1186/1471-2105-14-106>

Bleicher, D. (2023) *Building Regulations (TG 24/2023)*. Available at: https://www.bsria.com/uk/product/B6LgPn/building_regulations_tg_242023_a15d25e1/?_gl=1*1j4urze*_up*MQ..*_ga*NDQ1MzM5OTkuMTczMTAxMTIzMQ..*_ga_L0PN3DGN1B*MTczMTAxMTIzM4xLjAuMTczMTAxMTIzM4wLjAuMA.. (Accessed: 7 November 2024).

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence, Relevance 97, 245–271.
[https://doi.org/10.1016/S0004-3702\(97\)00063-5](https://doi.org/10.1016/S0004-3702(97)00063-5)

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis 143, 106839. <https://doi.org/10.1016/j.csda.2019.106839>

Boopathi, P., Gomathi, P., 2019. Scientometric analysis of diabetes research output during the year 2014-2018: Indexed by web of science. Library Philosophy and Practice.

Borhani, R., Borhani, S., Katsaggelos, A.K., 2022. Linear Regression, in: Borhani, R., Borhani, S., Katsaggelos, A.K. (Eds.), Fundamentals of Machine Learning and Deep Learning in Medicine. Springer International Publishing, Cham, pp. 69–87.
https://doi.org/10.1007/978-3-031-19502-0_4

Bornmann, L., Schier, H., Marx, W., Daniel, H.-D., 2012. What factors determine citation counts of publications in chemistry besides their quality? Journal of Informetrics 6, 11–18. <https://doi.org/10.1016/j.joi.2011.08.004>

Borowski, M., Zwolińska, K., 2020. Prediction of cooling energy consumption in hotel building using machine learning techniques. Energies 13.
<https://doi.org/10.3390/en13236226>

- Bouktif, S., Fiaz, A., Ouni, A., Serhani, M.A., 2020. Multi-Sequence LSTM-RNN Deep Learning and Metaheuristics for Electric Load Forecasting. *Energies* 13, 391. <https://doi.org/10.3390/en13020391>
- 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, 101533. <https://doi.org/10.1016/j.scs.2019.101533>
- 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 Appl. Sci.* 2, 297. <https://doi.org/10.1007/s42452-020-2024-9>
- Brannen, J., Coram, T., 1992. Mixing methods: Qualitative and quantitative research. Avebury Aldershot.
- BRE (2024) *BREEAM* from BRE, BRE Group. Available at: <https://bregroup.com/products/breeam> (Accessed: 7 November 2024).

Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., Utikal, Jochen S., von Kalle, C., Ludwig-Peitsch, W., Sirokay, J., Heinzerling, L., Albrecht, M., Baratella, K., Bischof, L., Chorti, E., Dith, A., Drusio, C., Giese, N., Gratsias, E., Griewank, K., Hallasch, S., Hanhart, Z., Herz, S., Hohaus, K., Jansen, P., Jockenhöfer, F., Kanaki, T., Knispel, S., Leonhard, K., Martaki, A., Matei, L., Matull, J., Olischewski, A., Petri, M., Placke, J.-M., Raub, S., Salva, K., Schlott, S., Sody, E., Steingrube, N., Stoffels, I., Uigurel, S., Zaremba, A., Gebhardt, C., Booken, N., Christolouka, M., Buder-Bakhaya, K., Bokor-Billmann, T., Enk, A., Gholam, P., Hänßle, H., Salzmann, M., Schäfer, S., Schäkel, K., Schank, T., Bohne, A.-S., Deffaa, S., Drerup, K., Egberts, F., Erkens, A.-S., Ewald, B., Falkvoll, S., Gerdes, S., Harde, V., Hauschild, A., Jost, M., Kosova, K., Messinger, L., Metzner, M., Morrison, K., Motamedi, R., Pinczker, A., Rosenthal, A., Scheller, N., Schwarz, T., Stölzl, D., Thielking, F., Tomaschewski, E., Wehkamp, U., Weichenthal, M., Wiedow, O., Bär, C.M., Bender-Säbelkampf, S., Horbrügger, M., Karoglan, A., Kraas, L., Faulhaber, J., Geraud, C., Guo, Z., Koch, P., Linke, M., Maurier, N., Müller, V., Thomas, B., Utikal, Jochen Sven, Alamri, A.S.M., Baczako, A., Berking, C., Betke, M., Haas, C., Hartmann, D., Heppt, M.V., Kilian, K., Krammer, S., Lapczynski, N.L., Mastnik, S., Nasifoglu, S., Ruini, C., Sattler, E., Schlaak, M., Wolff, H., Achatz, B., Bergbreiter, A., Drexler, K., Ettinger, M., Haferkamp, S., Halupczok, A., Hegemann, M., Dinauer, V., Maagk, M., Mickler, M.,

- Philipp, B., Wilm, A., Wittmann, C., Gesierich, A., Glutsch, V., Kahlert, K., Kerstan, A., Schilling, B., Schrüfer, P., 2019. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer* 113, 47–54. <https://doi.org/10.1016/j.ejca.2019.04.001>
- Bryman, A., 2003. *Quantity and Quality in Social Research*. Routledge.
- BSRIA (2024) *Building Services Research and Information Association (BSRIA)*. Available at: <https://www.bsria.com/uk/about/> (Accessed: 7 November 2024).
- Bulman, M., 2018. UK facing its biggest housing shortfall on record with backlog of 4 million homes, shows research [WWW Document]. The Independent. URL <https://www.independent.co.uk/news/uk/home-news/housing-homeless-crisis-homes-a8356646.html> (accessed 9.9.21).
- Burrell, G., Morgan, G., 1979. *Sociological Paradigms and Organisational Analysis: Elements of the Sociology of Corporate Life*. <https://doi.org/10.4324/9781315609751>
- Bustos, N., Tello, M.A., Dropelmann, G., García, N., Feijoo, F., Leiva, V., 2022. Machine learning techniques as an efficient alternative diagnostic tool for COVID-19 cases. *Signa Vitae* 18, 23–33. <https://doi.org/10.22514/sv.2021.110>
- Cai, M., Pipattanasomporn, M., Rahman, S., 2019. Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques. *Applied Energy* 236, 1078–1088. <https://doi.org/10.1016/j.apenergy.2018.12.042>
- Cajias, M., Piazolo, D., 2012. Green Performs Better: Energy Efficiency and Financial Return on Buildings. *Journal of Corporate Real Estate* 15, 53–72. <https://doi.org/10.1108/JCRE-12-2012-0031>
- Cao, L., Li, Y., Zhang, J., Jiang, Y., Han, Y., Wei, J., 2020. Electrical load prediction of healthcare buildings through single and ensemble learning. *Energy Reports* 6, 2751–2767. <https://doi.org/10.1016/j.egyr.2020.10.005>
- Capuano, D.L., 2019. International Energy Outlook 2020 (IEO2020) 7.
- Carrera, B., Peyrard, S., Kim, K., 2021. Meta-regression framework for energy consumption prediction in a smart city: A case study of Songdo in South Korea. *Sustainable Cities and Society* 72, 103025. <https://doi.org/10.1016/j.scs.2021.103025>
- Casebeer, A.L., Verhoef, M.J., 1997. Combining qualitative and quantitative research methods: Considering the possibilities for enhancing the study of chronic diseases. *Chronic diseases in Canada* 18, 130–135.

- Chae, Y.T., Horesh, R., Hwang, Y., Lee, Y.M., 2016. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy and Buildings* 111, 184–194. <https://doi.org/10.1016/j.enbuild.2015.11.045>
- Chammas, M., Makhoul, A., Demerjian, J., 2019. An efficient data model for energy prediction using wireless sensors. *Computers and Electrical Engineering* 76, 249–257. <https://doi.org/10.1016/j.compeleceng.2019.04.002>
- Chandrashekhar, G., Sahin, F., 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, 40th-year commemorative issue 40, 16–28. <https://doi.org/10.1016/j.compeleceng.2013.11.024>
- Chang, K.-H., 2012. Stochastic Nelder–Mead simplex method – A new globally convergent direct search method for simulation optimization. *European Journal of Operational Research* 220, 684–694. <https://doi.org/10.1016/j.ejor.2012.02.028>
- Chen, B., Liu, Q., Chen, H., Wang, L., Deng, T., Zhang, L., Wu, X., 2021. Multiobjective optimization of building energy consumption based on BIM-DB and LSSVM-NSGA-II. *Journal of Cleaner Production* 294. <https://doi.org/10.1016/j.jclepro.2021.126153>
- Chen, C., 2014. The citospace manual. *College of Computing and Informatics* 1, 1–84.
- Chen, H., Feng, Z., Cao, S.-J., 2021. Quantitative investigations on setting parameters of air conditioning (air-supply speed and temperature) in ventilated cooling rooms. *Indoor and Built Environment* 30, 99–113. <https://doi.org/10.1177/1420326X19887776>
- Chen, K., Chen, Kunlong, Wang, Q., He, Z., Hu, J., He, J., 2019. Short-Term Load Forecasting With Deep Residual Networks. *IEEE Transactions on Smart Grid* 10, 3943–3952. <https://doi.org/10.1109/TSG.2018.2844307>
- Chen, M., Zhang, W., Xie, L., Ni, Z., Wei, Q., Wang, W., Tian, H., 2019. Experimental and numerical evaluation of the crystalline silicon PV window under the climatic conditions in southwest China. *Energy* 183, 584–598. <https://doi.org/10.1016/j.energy.2019.06.146>
- Chen, Y., Guo, M., Chen, Zhisen, Chen, Zhe, 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, Y., Wang, S., Di, H., 2006. Study on the energy saving effect of residential windows. *Taiyangneng Xuebao/Acta Energiae Solaris Sinica* 27, 101–105.
- Chen, Y., Zhang, F., Berardi, U., 2020. Day-ahead prediction of hourly subentry energy consumption in the building sector using pattern recognition algorithms. *Energy* 211. <https://doi.org/10.1016/j.energy.2020.118530>

- Chen, Y.-T., Piedad, E., Kuo, C.-C., 2019. Energy Consumption Load Forecasting Using a Level-Based Random Forest Classifier. *Symmetry* 11, 956.
<https://doi.org/10.3390/sym11080956>
- Cherven, K., 2015. Mastering Gephi network visualization. Packt Publishing Ltd.
- Chiesa, G., Acquaviva, A., Grossi, M., Bottaccioli, L., Floridia, M., Pristeri, E., Sanna, E., 2019. Parametric Optimization of Window-to-Wall Ratio for Passive Buildings Adopting A Scripting Methodology to Dynamic-Energy Simulation. *Sustainability* 11, 3078. <https://doi.org/10.3390/su11113078>
- Chiradeja, P., Ngaopitakkul, A., 2019. Energy and economic analysis of tropical building envelope material in compliance with Thailand's building energy code. *Sustainability (Switzerland)* 11. <https://doi.org/10.3390/su11236872>
- Chirarattananon, S., Taveekun, J., 2004. An OTTV-based energy estimation model for commercial buildings in Thailand. *Energy and Buildings, Building Research and the Sustainability of the Built Environment in the Tropics* 36, 680–689.
<https://doi.org/10.1016/j.enbuild.2004.01.035>
- Cho, S., Lee, J., Baek, J., Kim, G.-S., Leigh, S.-B., 2019. Investigating primary factors affecting electricity consumption in non-residential buildings using a data-driven approach. *Energies* 12. <https://doi.org/10.3390/en12214046>
- Cho, Sooyoun, Lee, J., Baek, J., Kim, G.-S., Leigh, S.-B., 2019. Investigating Primary Factors Affecting Electricity Consumption in Non-Residential Buildings Using a Data-Driven Approach. *Energies* 12, 4046. <https://doi.org/10.3390/en12214046>
- Chokwitthaya, C., Zhu, Y., Dibiano, R., Mukhopadhyay, S., 2020. A machine learning algorithm to improve building performance modeling during design. *MethodsX* 7, 100726. <https://doi.org/10.1016/j.mex.2019.10.037>
- Chou, J.-S., Bui, D.-K., 2014. Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Buildings* 82, 437–446.
<https://doi.org/10.1016/j.enbuild.2014.07.036>
- Chou, J.-S., Tran, D.-S., 2018. Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. *Energy* 165, 709–726. <https://doi.org/10.1016/j.energy.2018.09.144>
- Chung, K.C., Tan, S.S., Holdsworth, D.K., 2008. Insolvency Prediction Model Using Multivariate Discriminant Analysis and Artificial Neural Network for the Finance Industry in New Zealand (SSRN Scholarly Paper No. ID 1080430). Social Science Research Network, Rochester, NY.

CIBSE (2024). Available at: <https://www.cibse.org/about-cibse> (Accessed: 7 November 2024).

- Çiftsüren, M.N., Akkol, S., 2018. Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, LASSO and elastic net. Archives Animal Breeding 61, 279–284. <https://doi.org/10.5194/aab-61-279-2018>
- Ciulla, G., D'Amico, A., 2019. Building energy performance forecasting: A multiple linear regression approach. Applied Energy 253. <https://doi.org/10.1016/j.apenergy.2019.113500>
- Ciulla, G., Lo Brano, V., D'Amico, A., 2016. Modelling relationship among energy demand, climate and office building features: A cluster analysis at European level. Applied Energy 183, 1021–1034. <https://doi.org/10.1016/j.apenergy.2016.09.046>
- Climate change, 2011. . Nature 479, 267–268. <https://doi.org/10.1038/479267b>
- Cohen, L., Manion, L., Morrison, K., 2007. Research methods in education, 6th ed. ed. Routledge, London ; New York.
- Collis, J., Hussey, R., 2009. Business Research: A Practical Guide for Undergraduate and Postgraduate Students, 3rd edition. ed. Palgrave Macmillan, Hampshire, UK ; New York, NY.
- Collis, J., Hussey, R., 2003. Business Research. Palgrave Macmillan.
- Colmenar-Santos, A., Terán de Lober, L.N., Borge-Diez, D., Castro-Gil, M., 2013. Solutions to reduce energy consumption in the management of large buildings. Energy and Buildings 56, 66–77. <https://doi.org/10.1016/j.enbuild.2012.10.004>
- Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Cozby, P., Bates, S., 2012. Methods in Behavioral Research.
- Crawley, D.B., Hand, J.W., Kummert, M., Griffith, B.T., 2008. Contrasting the capabilities of building energy performance simulation programs. Building and Environment, Part Special: Building Performance Simulation 43, 661–673. <https://doi.org/10.1016/j.buildenv.2006.10.027>
- Culaba, A.B., Del Rosario, A.J.R., Ubando, A.T., Chang, J.-S., 2020. Machine learning-based energy consumption clustering and forecasting for mixed-use buildings. International Journal of Energy Research 44, 9659–9673. <https://doi.org/10.1002/er.5523>

- Curtis, J., Devitt, N., Whelan, A., 2014. Estimating Building Energy Ratings for the Residential Building Stock: Location and Occupancy (No. WP489), Papers, Papers. Economic and Social Research Institute (ESRI).
- da Silva, D.G., Geller, M.T.B., dos Santos Moura, M.S., de Moura Meneses, A.A., 2022. Performance evaluation of LSTM neural networks for consumption prediction. e-Prime-Advances in Electrical Engineering, Electronics and Energy 2, 100030.
- Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.M., Adetunmbi, A.O., Ajibawa, O.E., 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* 5, e01802. <https://doi.org/10.1016/j.heliyon.2019.e01802>
- Dalal, K.R., 2018. Review on Application of Machine learning Algorithm for Data Science, in: 2018 3rd International Conference on Inventive Computation Technologies (ICICT). IEEE, pp. 270–273.
- D'Amico, A., Ciulla, G., Traverso, M., Lo Brano, V., Palumbo, E., 2019. Artificial Neural Networks to assess energy and environmental performance of buildings: An Italian case study. *Journal of Cleaner Production* 239.
<https://doi.org/10.1016/j.jclepro.2019.117993>
- Dandotiya, B., 2020. Climate-Change-and-Its-Impact-on-Terrestrial-Ecosystems.
<https://doi.org/10.4018/978-1-7998-3343-7.ch007>
- Darko, A., Chan, A.P.C., Adabre, M.A., Edwards, D.J., Hosseini, M.R., Ameyaw, E.E., 2020. Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. *Automation in Construction* 112, 103081.
<https://doi.org/10.1016/j.autcon.2020.103081>
- Dash, M., Liu, H., 1997. Feature selection for classification. *Intelligent Data Analysis* 1, 131–156. [https://doi.org/10.1016/S1088-467X\(97\)00008-5](https://doi.org/10.1016/S1088-467X(97)00008-5)
- Deb, C., Zhang, F., Yang, J., Lee, S.E., Shah, K.W., 2017a. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews* 74, 902–924. <https://doi.org/10.1016/j.rser.2017.02.085>
- Deb, C., Zhang, F., Yang, J., Lee, S.E., Shah, K.W., 2017b. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews* 74, 902–924. <https://doi.org/10.1016/j.rser.2017.02.085>
- Debrah, C., Chan, A.P.C., Darko, A., 2022. Artificial intelligence in green building. *Automation in Construction* 137, 104192.
<https://doi.org/10.1016/j.autcon.2022.104192>

- Demertzis, K., Kostinakis, K., Morfidis, K., Iliadis, L., 2022. A Comparative Evaluation of Machine Learning Algorithms for the Prediction of R/C Buildings' Seismic Damage.
- Demsar, J., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets 30.
- Department of Energy and Climate Change, 2009. Impact Assessment of the Climate Change Act.
- Diebold, F.X., 2012. On the Origin(s) and Development of the Term “Big Data.”
<https://doi.org/10.2139/ssrn.2152421>
- Dietterich, T.G., 2000. Ensemble Methods in Machine Learning, in: Multiple Classifier Systems, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 1–15.
https://doi.org/10.1007/3-540-45014-9_1
- Diirr, B., Santos, G., 2019. Interorganizational Information Systems: Systematic Literature Mapping Protocol. RelaTe-DIA.
- Ding, H., Feng, P.-M., Chen, W., Lin, H., 2014. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol. BioSyst. 10, 2229–2235.
<https://doi.org/10.1039/C4MB00316K>
- Ding, Y., Liu, X., 2020. A comparative analysis of data-driven methods in building energy benchmarking. Energy and Buildings 209, 109711.
<https://doi.org/10.1016/j.enbuild.2019.109711>
- Ding, Z., Chen, W., Hu, T., Xu, X., 2021. Evolutionary double attention-based long short-term memory model for building energy prediction: Case study of a green building. Applied Energy 288, 116660. <https://doi.org/10.1016/j.apenergy.2021.116660>
- Ding, Z., Zhu, M., Tam, V.W.Y., Yi, G., Tran, C.N.N., 2018. A system dynamics-based environmental benefit assessment model of construction waste reduction management at the design and construction stages. Journal of Cleaner Production 176, 676–692.
<https://doi.org/10.1016/j.jclepro.2017.12.101>
- Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., Torres, J.F., 2018a. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. Energies 11, 949. <https://doi.org/10.3390/en11040949>
- Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., Torres, J.F., 2018b. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. Energies 11, 949. <https://doi.org/10.3390/en11040949>
- Divina, F., Torres, M.G., Vela, F.A.G., Noguera, J.L.V., 2019. A comparative study of time series forecasting methods for short term electric energy consumption prediction in smart buildings. Energies 12. <https://doi.org/10.3390/en12101934>

- Domingos, P., 2012. A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. <https://doi.org/10.1145/2347736.2347755>
- Dong, B., Cao, C., Lee, S.E., 2005. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings* 37, 545–553. <https://doi.org/10.1016/j.enbuild.2004.09.009>
- Dong, B., Li, Z., Rahman, S.M.M., Vega, R., 2016. A hybrid model approach for forecasting future residential electricity consumption. *Energy and Buildings* 117, 341–351. <https://doi.org/10.1016/j.enbuild.2015.09.033>
- Dong, Z., Liu, J., Liu, B., Li, K., Li, X., 2021a. 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>
- Dong, Z., Liu, J., Liu, B., Li, K., Li, X., 2021b. 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>
- Donoghue, J.O., Roantree, M., 2015. A Framework for Selecting Deep Learning Hyper-parameters, in: Maneth, S. (Ed.), *Data Science, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 120–132. https://doi.org/10.1007/978-3-319-20424-6_12
- Dun, M., Wu, L., 2020. Forecasting the Building Energy Consumption in China Using Grey Model. *Environmental Processes* 7, 1009–1022. <https://doi.org/10.1007/s40710-020-00438-3>
- Easterby-Smith, M., Thorpe, R., Jackson, P., Lowe, A., 2008. *Management Research*. SAGE.
- Easterby-Smith, M., Thorpe, R., Lowe, A., 2001. *Management Research: An Introduction*, Second edition. ed. SAGE Publications Ltd, London.
- Effrosynidis, D., Arampatzis, A., 2021. An evaluation of feature selection methods for environmental data. *Ecological Informatics* 61, 101224. <https://doi.org/10.1016/j.ecoinf.2021.101224>
- Egwim, C.N., Alaka, H., Toriola-Coker, L.O., Balogun, H., Sunmola, F., 2021. Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications* 6, 100166. <https://doi.org/10.1016/j.mlwa.2021.100166>
- EIA, 2020. *Monthly Energy Review – October 2020* 272.

- Eseye, A.T., Lehtonen, M., 2020. Short-Term Forecasting of Heat Demand of Buildings for Efficient and Optimal Energy Management Based on Integrated Machine Learning Models. *IEEE Transactions on Industrial Informatics* 16, 7743–7755.
<https://doi.org/10.1109/TII.2020.2970165>
- European Parliament, 2002. Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings (repealed) [WWW Document]. <https://webarchive.nationalarchives.gov.uk/eu-exit/https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02002L0091-20081211>. URL <https://www.legislation.gov.uk/eudr/2002/91/2008-12-11> (accessed 6.15.21).
- Faisal, H.M., Javaid, N., Sarfraz, B., Baqi, A., Bilal, M., Haider, I., Shuja, S.M., 2019. Prediction of Building Energy Consumption Using Enhanced Convolutional Neural Network, in: Barolli, L., Takizawa, M., Xhafa, F., Enokido, T. (Eds.), Web, Artificial Intelligence and Network Applications, Advances in Intelligent Systems and Computing. Springer International Publishing, Cham, pp. 1157–1168.
https://doi.org/10.1007/978-3-030-15035-8_111
- Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G., 2008. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal* 22, 338–342. <https://doi.org/10.1096/fj.07-9492LSF>
- Fan, C., Sun, Y., Xiao, F., Ma, J., Lee, D., Wang, J., Tseng, Y.C., 2020. Statistical investigations of transfer learning-based methodology for short-term building energy predictions. *Applied Energy* 262. <https://doi.org/10.1016/j.apenergy.2020.114499>
- Fan, C., Xiao, F., Wang, S., 2014a. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy* 127, 1–10. <https://doi.org/10.1016/j.apenergy.2014.04.016>
- Fan, C., Xiao, F., Wang, S., 2014b. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy* 127, 1–10.
- Fan, Cheng, Xiao, F., Zhao, Y., 2017. A short-term building cooling load prediction method using deep learning algorithms. *Applied Energy* 195, 222–233.
<https://doi.org/10.1016/j.apenergy.2017.03.064>
- Fan, C., Xiao, F., Zhao, Y., 2017. A short-term building cooling load prediction method using deep learning algorithms. *Applied Energy* 195, 222–233.
<https://doi.org/10.1016/j.apenergy.2017.03.064>

- Fan, X., 2022. A method for the generation of typical meteorological year data using ensemble empirical mode decomposition for different climates of China and performance comparison analysis. *Energy* 240.
<https://doi.org/10.1016/j.energy.2021.122822>
- Fang, X., Gong, G., Li, G., Chun, L., Li, W., Peng, P., 2021. A hybrid deep transfer learning strategy for short term cross-building energy prediction. *Energy* 215.
<https://doi.org/10.1016/j.energy.2020.119208>
- Fathi, Soheil, Srinivasan, R., Fenner, A., Fathi, Sahand, 2020a. Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews* 133, 110287.
<https://doi.org/10.1016/j.rser.2020.110287>
- Fathi, Soheil, Srinivasan, R., Fenner, A., Fathi, Sahand, 2020b. Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews* 133, 110287.
<https://doi.org/10.1016/j.rser.2020.110287>
- Feng, C., Zhang, J., 2020. Assessment of aggregation strategies for machine-learning based short-term load forecasting. *Electric Power Systems Research* 184.
<https://doi.org/10.1016/j.epsr.2020.106304>
- Fernandez-Antolin, M.-M., del Río, J.M., Costanzo, V., Nocera, F., Gonzalez-Lezcano, R.-A., 2019. Passive design strategies for residential buildings in different Spanish climate zones. *Sustainability (Switzerland)* 11. <https://doi.org/10.3390/su11184816>
- Feurer, M., Hutter, F., 2019. Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges* 3–33.
- Flores, V., Keith, B., 2019. Gradient Boosted Trees Predictive Models for Surface Roughness in High-Speed Milling in the Steel and Aluminum Metalworking Industry. *Complexity* 2019, e1536716. <https://doi.org/10.1155/2019/1536716>
- Florides, G.A., Tassou, S.A., Kalogirou, S.A., Wrobel, L.C., 2002. Measures used to lower building energy consumption and their cost effectiveness. *Applied Energy* 73, 299–328. [https://doi.org/10.1016/S0306-2619\(02\)00119-8](https://doi.org/10.1016/S0306-2619(02)00119-8)
- Fumo, N., Rafe Biswas, M.A., 2015. Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews* 47, 332–343.
<https://doi.org/10.1016/j.rser.2015.03.035>

- Gagliano, A., Detommaso, M., Nocera, F., Evola, G., 2015. A multi-criteria methodology for comparing the energy and environmental behavior of cool, green and traditional roofs. *Building and Environment* 90, 71–81. <https://doi.org/10.1016/j.buildenv.2015.02.043>
- Gao, W., Alsarraf, J., Moayedi, H., Shahsavar, A., Nguyen, H., 2019. Comprehensive preference learning and feature validity for designing energy-efficient residential buildings using machine learning paradigms. *Applied Soft Computing* 84, 105748. <https://doi.org/10.1016/j.asoc.2019.105748>
- Gao, X., Qi, C., Xue, G., Song, J., Zhang, Y., Yu, S., 2020. Forecasting the Heat Load of Residential Buildings with Heat Metering Based on CEEMDAN-SVR. *Energies* 13, 6079. <https://doi.org/10.3390/en13226079>
- Geraldi, M.S., Ghisi, E., 2022. Integrating evidence-based thermal satisfaction in energy benchmarking: A data-driven approach for a whole-building evaluation. *Energy* 244. <https://doi.org/10.1016/j.energy.2022.123161>
- Ghosh, A., Neogi, S., 2018. Effect of fenestration geometrical factors on building energy consumption and performance evaluation of a new external solar shading device in warm and humid climatic condition. *Solar Energy* 169, 94–104. <https://doi.org/10.1016/j.solener.2018.04.025>
- Gonzalez-Abril, L., Nuñez, H., Angulo, C., Velasco, F., 2014. GSVM: An SVM for handling imbalanced accuracy between classes in bi-classification problems. *Applied Soft Computing* 17, 23–31. <https://doi.org/10.1016/j.asoc.2013.12.013>
- Gopi, E.S., 2020. Regression Techniques, in: Gopi, E.S. (Ed.), *Pattern Recognition and Computational Intelligence Techniques Using Matlab*. Springer International Publishing, Cham, pp. 69–120. https://doi.org/10.1007/978-3-030-22273-4_3
- Goyal, K., Tiwari, N., Sonekar, J., 2020. An Anatomization of Data Classification Based on Machine Learning Techniques. *IJRAR-International Journal of Research and Analytical Reviews (IJRAR)* 7, 713–716.
- Grix, J., 2004. *The Foundations of Research*, 2nd Edition. ed. Palgrave, Basingstoke.
- Gros, A., Bozonnet, E., Inard, C., 2014. Cool materials impact at district scale - Coupling building energy and microclimate models. *Sustainable Cities and Society* 13, 254–266. <https://doi.org/10.1016/j.scs.2014.02.002>
- Groß, A., Lenders, A., Schwenker, F., Braun, D.A., Fischer, D., 2021. Comparison of short-term electrical load forecasting methods for different building types. *Energy Informatics* 4. <https://doi.org/10.1186/s42162-021-00172-6>

- Gul, M.S., NezamiFar, E., 2020. Investigating the interrelationships among occupant attitude, knowledge and behaviour in LEED-certified buildings using structural equation modelling. *Energies* 13. <https://doi.org/10.3390/en13123158>
- Gunantara, N., 2018. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering* 5, 1502242. <https://doi.org/10.1080/23311916.2018.1502242>
- Guo, N., Gui, W., Chen, W., Tian, X., Qiu, W., Tian, Z., Zhang, X., 2020a. Using improved support vector regression to predict the transmitted energy consumption data by distributed wireless sensor network. *J Wireless Com Network* 2020, 120. <https://doi.org/10.1186/s13638-020-01729-x>
- Guo, N., Gui, W., Chen, W., Tian, X., Qiu, W., Tian, Z., Zhang, X., 2020b. Using improved support vector regression to predict the transmitted energy consumption data by distributed wireless sensor network. *EURASIP Journal on Wireless Communications and Networking* 2020, 120. <https://doi.org/10.1186/s13638-020-01729-x>
- Guo, Y., Wang, J., Chen, H., Li, G., Liu, J., Xu, C., Huang, R., Huang, Y., 2018. Machine learning-based thermal response time ahead energy demand prediction for building heating systems. *Applied Energy* 221, 16–27. <https://doi.org/10.1016/j.apenergy.2018.03.125>
- Hamed, M., Nada, S., 2019. Statistical Analysis for Economics of the Energy Development in North Zone of Cairo. *International Journal of Finance & Economics* 5, 140–160.
- Hao, J., Ho, T.K., 2019. Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics* 44, 348–361. <https://doi.org/10.3102/1076998619832248>
- Hasan, O.A., Defer, D., 2019. The role of new technologies in understanding the building energy performance: a comparative study. *International Journal of Smart Grid and Clean Energy* 8, 397–401. <https://doi.org/10.12720/sgce.8.4.397-401>
- Hekler, A., Utikal, J.S., Enk, A.H., Solass, W., Schmitt, M., Klode, J., Schadendorf, D., Sondermann, W., Franklin, C., Bestvater, F., Flaig, M.J., Krahl, D., von Kalle, C., Fröhling, S., Brinker, T.J., 2019. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer* 118, 91–96. <https://doi.org/10.1016/j.ejca.2019.06.012>
- Henderson, G. and Shorrock, L.D. (1986) ‘BREDEM—The BRE domestic energy model: Testing the predictions of a two-zone version’, *Building Services Engineering Research and Technology*, 7(2), pp. 87–91. Available at: <https://doi.org/10.1177/014362448600700205>.

- Himeur, Y., Alsalemi, A., Bensaali, F., Amira, A., 2020. Building power consumption datasets: Survey, taxonomy and future directions. *Energy and Buildings* 227, 110404. <https://doi.org/10.1016/j.enbuild.2020.110404>
- Hoang, D.T., Kang, H.J., 2019. Rotary Machine Fault Diagnosis Using Scalogram Image and Convolutional Neural Network with Batch Normalization, in: Huang, D.-S., Huang, Z.-K., Hussain, A. (Eds.), *Intelligent Computing Methodologies, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 283–293. https://doi.org/10.1007/978-3-030-26766-7_26
- Hosseini, S., Fard, R.H., 2021. Machine Learning Algorithms for Predicting Electricity Consumption of Buildings. *Wireless Personal Communications* 121, 3329–3341. <https://doi.org/10.1007/s11277-021-08879-1>
- Hribar, R., Potočnik, P., Šilc, J., Papa, G., 2019. A comparison of models for forecasting the residential natural gas demand of an urban area. *Energy* 167, 511–522. <https://doi.org/10.1016/j.energy.2018.10.175>
- Hsu, D., 2015. Identifying key variables and interactions in statistical models of building energy consumption using regularization. *Energy* 83, 144–155. <https://doi.org/10.1016/j.energy.2015.02.008>
- Huang, Y., Niu, J.-L., Chung, T.-M., 2013. Study on performance of energy-efficient retrofitting measures on commercial building external walls in cooling-dominant cities. *Applied Energy* 103, 97–108. <https://doi.org/10.1016/j.apenergy.2012.09.003>
- Huljanah, M., Rustam, Z., Utama, S., Siswantining, T., 2019. Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. *IOP Conf. Ser.: Mater. Sci. Eng.* 546, 052031. <https://doi.org/10.1088/1757-899X/546/5/052031>
- Hung, Y.-N., Yang, Y.-H., 2018. Frame-level Instrument Recognition by Timbre and Pitch. [arXiv:1806.09587 \[cs, eess\]](https://arxiv.org/abs/1806.09587).
- Hwang, J., Suh, D., Otto, M.-O., 2020. Forecasting electricity consumption in commercial buildings using a machine learning approach. *Energies* 13. <https://doi.org/10.3390/en13225885>
- Ibraheem, T.B., Salmanu, H., Bashir, T.S., Adamu, H.S., 2017. Renewable Energy Integration in African Buildings: Criteria and Prospects. *American Journal of Engineering Research (AJER)* Volume-6, pp-39-43.

- Ibrahim, I.A., Khatib, T., 2017. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm.
<https://doi.org/10.1016/J.ENCONMAN.2017.02.006>
- Ifeoma, A.J., Akande, I., 2021. Impact of Building Orientation on Building Performance 6, 16.
- Ihara, T., Gustavsen, A., Jelle, B.P., 2015. Effect of facade components on energy efficiency in office buildings. Applied Energy 158, 422–432.
<https://doi.org/10.1016/j.apenergy.2015.08.074>
- Ihsan, Y.N., 2020. Marine macro debris transport based on hydrodynamic model before and after reclamation in Jakarta Bay, Indonesia. Malaysian Journal of Applied Sciences 5, 100–111. <https://doi.org/10.37231/myjas.2020.5.2.241>
- Iken, O., Dlimi, M., Agounoun, R., Kadiri, I., Fertahi, S.E.-D., Zoubir, A., Sbai, K., 2019. Numerical investigation of energy performance and cost analysis of Moroccan's building smart walls integrating vanadium dioxide. Solar Energy 179, 249–263.
<https://doi.org/10.1016/j.solener.2018.12.062>
- Ilager, S., Ramamohanarao, K., Buyya, R., 2021. Thermal Prediction for Efficient Energy Management of Clouds Using Machine Learning. IEEE Transactions on Parallel and Distributed Systems 32, 1044–1056. <https://doi.org/10.1109/TPDS.2020.3040800>
- International Energy Agency: Cooling [WWW Document], 2019. . IEA. URL
<https://www.iea.org/reports/cooling> (accessed 7.28.22).
- Iqbal, M., Muneeb Abid, M., Noman, M., Manzoor, Engr.Dr.A., 2020. Review of feature selection methods for text classification. International Journal of Advanced Computer Research 10, 2277–7970. <https://doi.org/10.19101/IJACR.2020.1048037>
- Islam, M.Z., Moore, R., Cosco, N., 2016. Child-Friendly, Active, Healthy Neighborhoods: Physical Characteristics and Children's Time Outdoors. Environment and Behavior 48, 711–736. <https://doi.org/10.1177/0013916514554694>
- Ivanov, T., Korfiatis, N., Zicari, R., 2013. On the inequality of the 3V's of Big Data Architectural Paradigms: A case for heterogeneity. ArXiv Prepr.
- Izidio, Diogo M. F., de Mattos Neto, P.S.G., Barbosa, L., de Oliveira, J.F.L., Marinho, M.H. da N., Rissi, G.F., 2021. Evolutionary Hybrid System for Energy Consumption Forecasting for Smart Meters. Energies 14, 1794. <https://doi.org/10.3390/en14071794>
- Izidio, D.M.F., de Mattos Neto, P.S.G., Barbosa, L., de Oliveira, J.F.L., Marinho, M.H.D.N., Rissi, G.F., 2021. Evolutionary hybrid system for energy consumption forecasting for smart meters. Energies 14. <https://doi.org/10.3390/en14071794>

- Jaber, S., Ajib, S., 2011. Optimum, technical and energy efficiency design of residential building in Mediterranean region. *Energy and Buildings* 43, 1829–1834.
<https://doi.org/10.1016/j.enbuild.2011.03.024>
- Jahromi, A.H., Taheri, M., 2017. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features, in: 2017 Artificial Intelligence and Signal Processing Conference (AISP). Presented at the 2017 Artificial Intelligence and Signal Processing Conference (AISP), pp. 209–212.
<https://doi.org/10.1109/AISP.2017.8324083>
- Jang, J., Han, J., Kim, M.-H., Kim, D.-W., Leigh, S.-B., 2021. Extracting influential factors for building energy consumption via data mining approaches. *Energies* 14.
<https://doi.org/10.3390/en14248505>
- Jang, J., Lee, J., Son, E., Park, K., Kim, G., Lee, J.H., Leigh, S.-B., 2019. Development of an improved model to predict building thermal energy consumption by utilizing feature selection. *Energies* 12. <https://doi.org/10.3390/en12214187>
- Jhai, 2023. . Jhai. URL <https://jhai.co.uk/> (accessed 6.6.23).
- Jin, X., Xu, A., Bie, R., Guo, P., 2006. Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles, in: Li, J., Yang, Q., Tan, A.-H. (Eds.), Data Mining for Biomedical Applications, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 106–115.
https://doi.org/10.1007/11691730_11
- Jin, Y., Yan, D., Kang, X., Chong, A., Sun, H.-, Zhan, S., 2021. Forecasting building occupancy: A temporal-sequential analysis and machine learning integrated approach. *Energy and Buildings* 252, 111362. <https://doi.org/10.1016/j.enbuild.2021.111362>
- Jing, W., Zhen, M., Guan, H., Luo, W., Liu, X., 2022. A prediction model for building energy consumption in a shopping mall based on Chaos theory. *Energy Reports* 8, 5305–5312. <https://doi.org/10.1016/j.egyr.2022.03.205>
- Johnson, J., DelGiudice, B., Bangari, D.S., Peterson, E., Ulinski, G., Ryan, S., Thurberg, B.L., 2019. Lung (Inflated), in: The Laboratory Mouse. CRC Press.
- Johnson, R., Onwuegbuzie, A., 2004. Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational researcher* 33, 14.
<https://doi.org/10.3102/0013189X033007014>
- Jović, A., Brkić, K., Bogunović, N., 2015. A review of feature selection methods with applications, in: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Presented

- at the 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200–1205.
<https://doi.org/10.1109/MIPRO.2015.7160458>
- Kabir, M.A., 2020. Vehicle Speed Prediction based on Road Status using Machine Learning. Advanced Research in Energy and Engineering 2.
- Kadkhodaei, H.R., Moghadam, A.M.E., Dehghan, M., 2020. HBoost: A heterogeneous ensemble classifier based on the Boosting method and entropy measurement. Expert Systems with Applications 157, 113482. <https://doi.org/10.1016/j.eswa.2020.113482>
- Kamel, E., Sheikh, S., Huang, X., 2020. Data-driven predictive models for residential building energy use based on the segregation of heating and cooling days. Energy 206. <https://doi.org/10.1016/j.energy.2020.118045>
- Kanyongo, W., Ezugwu, A., 2023. Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. Informatics in Medicine Unlocked 38, 101232.
<https://doi.org/10.1016/j imu.2023.101232>
- Kapetanakis, D.-S., Mangina, E., Finn, D.P., 2017. Input variable selection for thermal load predictive models of commercial buildings. Energy and Buildings 137, 13–26.
<https://doi.org/10.1016/j.enbuild.2016.12.016>
- Karatasou, S., Santamouris, M., Geros, V., 2006. Modeling and predicting building's energy use with artificial neural networks: Methods and results. Energy and Buildings 38, 949–958. <https://doi.org/10.1016/j.enbuild.2005.11.005>
- Karatzas, S.K., Chassiakos, A.P., Karameros, A.I., 2020. Business processes and comfort demand for energy flexibility analysis in buildings. Energies 13.
<https://doi.org/10.3390/en13246561>
- Kaur, K., Gupta, O.P., 2017. A machine learning approach to determine maturity stages of tomatoes. Oriental journal of computer science and technology 10, 683–690.
- Khan, A.-N., Iqbal, N., Rizwan, A., Ahmad, R., Kim, D.-H., 2021. An ensemble energy consumption forecasting model based on spatial-temporal clustering analysis in residential buildings. Energies 14. <https://doi.org/10.3390/en14113020>
- Khan, P.W., Byun, Y.-C., Lee, S.-J., Kang, D.-H., Kang, J.-Y., Park, H.-S., 2020. Machine Learning-Based Approach to Predict Energy Consumption of Renewable and Nonrenewable Power Sources. Energies 13, 4870.
<https://doi.org/10.3390/en13184870>

- Khantach, A.E., Hamlich, M., Belbounaguia, N. eddine, Khantach, A.E., Hamlich, M., Belbounaguia, N. eddine, 2019. Short-term load forecasting using machine learning and periodicity decomposition. *AIMS Energy* 7, 382–394.
<https://doi.org/10.3934/energy.2019.3.382>
- Khosravani, H., Castilla, M.D.M., Berenguel, M., Ruano, A., Ferreira, P., 2016. A Comparison of Energy Consumption Prediction Models Based on Neural Networks of a Bioclimatic Building. <https://doi.org/10.3390/EN9010057>
- Kim, D.D., Suh, H.S., 2021. Heating and cooling energy consumption prediction model for high-rise apartment buildings considering design parameters. *Energy for Sustainable Development* 61, 1–14. <https://doi.org/10.1016/j.esd.2021.01.001>
- Kim, J.-H., Seong, N.-C., Choi, W., 2020. Forecasting the Energy Consumption of an Actual Air Handling Unit and Absorption Chiller Using ANN Models. *Energies* 13, 4361.
<https://doi.org/10.3390/en13174361>
- Kim, M.K., Kim, Y.-S., Srebric, J., 2020. Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. *Sustainable Cities and Society* 62.
<https://doi.org/10.1016/j.scs.2020.102385>
- Kim, Moon Keun, Kim, Y.-S., Srebric, J., 2020. Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. *Sustainable Cities and Society* 62, 102385. <https://doi.org/10.1016/j.scs.2020.102385>
- Kim, Y.K., Bande, L., Aoul, K.A.T., Altan, H., 2021. Dynamic energy performance gap analysis of a university building: Case studies at uae university campus, UAE. *Sustainability (Switzerland)* 13, 1–15. <https://doi.org/10.3390/su13010120>
- Kira, K., Rendell, L.A., 1992. A Practical Approach to Feature Selection, in: Sleeman, D., Edwards, P. (Eds.), *Machine Learning Proceedings 1992*. Morgan Kaufmann, San Francisco (CA), pp. 249–256. <https://doi.org/10.1016/B978-1-55860-247-2.50037-1>
- Kitchin, R., McArdle, G., 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3, 2053951716631130.
<https://doi.org/10.1177/2053951716631130>
- Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. *Artificial Intelligence, Relevance* 97, 273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X)

- Köhler, N.D., Büttner, M., Andriamanga, N., Theis, F.J., 2021. Deep learning does not outperform classical machine learning for cell-type annotation.
<https://doi.org/10.1101/653907>
- Kolaitis, D.I., Malliotakis, E., Kontogeorgos, D.A., Mandilaras, I., Katsourinis, D.I., Founti, M.A., 2013. Comparative assessment of internal and external thermal insulation systems for energy efficient retrofitting of residential buildings. *Energy and Buildings* 64, 123–131. <https://doi.org/10.1016/j.enbuild.2013.04.004>
- Kolter, J., Ferreira, J., 2011. A Large-Scale Study on Predicting and Contextualizing Building Energy Usage. Proceedings of the AAAI Conference on Artificial Intelligence 25, 1349–1356.
- Konasani, V.R., Kadre, S., 2015. Logistic Regression, in: Konasani, V.R., Kadre, S. (Eds.), Practical Business Analytics Using SAS: A Hands-on Guide. Apress, Berkeley, CA, pp. 401–440. https://doi.org/10.1007/978-1-4842-0043-8_11
- Kong, W., Dong, Z.Y., Jia, Y., Hill, D.J., Xu, Y., Zhang, Y., 2019. Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network. *IEEE Transactions on Smart Grid* 10, 841–851. <https://doi.org/10.1109/TSG.2017.2753802>
- Kontokosta, C.E., Tull, C., 2017. A data-driven predictive model of city-scale energy use in buildings. *Applied Energy* 197, 303–317.
<https://doi.org/10.1016/j.apenergy.2017.04.005>
- Košir, M., Pajek, L., Iglič, N., Kunič, R., 2018. A theoretical study on a coupled effect of building envelope solar properties and thermal transmittance on the thermal response of an office cell. *Solar Energy* 174, 669–682.
<https://doi.org/10.1016/j.solener.2018.09.042>
- Kotsiantis, S.B., Kanellopoulos, D., Pintelas, P.E., 2006. Data preprocessing for supervised learning. *International journal of computer science* 1, 111–117.
- Koukaras, P., Bezas, N., Gkaidatzis, P., Ioannidis, D., Tzovaras, D., Tjortjis, C., 2021. Introducing a novel approach in one-step ahead energy load forecasting. *Sustainable Computing: Informatics and Systems* 32.
<https://doi.org/10.1016/j.suscom.2021.100616>
- Krarti, M., 2003. An Overview of Artificial Intelligence-Based Methods for Building Energy Systems. *Journal of Solar Energy Engineering* 125, 331–342.
<https://doi.org/10.1115/1.1592186>
- Krishnamoorthy, G., Ramakrishnan, J., Devi, S., 2009. Bibliometric analysis of literature on diabetes (1995 – 2004) 6.

- Krishnamurthi, K., Thapa, S., Kothari, L., Prakash, A., 2015. Arduino based weather monitoring system. International Journal of Engineering Research and General Science 3, 452–458.
- Kumar, P., Sinha, K., Nere, N.K., Shin, Y., Ho, R., Mlinar, L.B., Sheikh, A.Y., 2020. A machine learning framework for computationally expensive transient models. Sci Rep 10, 11492. <https://doi.org/10.1038/s41598-020-67546-w>
- Kumar, V., 2014. Feature Selection: A literature Review. SmartCR 4. <https://doi.org/10.6029/smartercr.2014.03.007>
- Kunasekaran, K.K.H., Sugumaran, R., 2016. Exploratory Analysis of Feature Selection Techniques in Medical Image Processing. Medical Image Processing 5.
- Kusiak, A., Li, M., Zhang, Z., 2010. A data-driven approach for steam load prediction in buildings. Applied Energy 87, 925–933. <https://doi.org/10.1016/j.apenergy.2009.09.004>
- Kuster, C., Rezgui, Y., Mourshed, M., 2017. Electrical load forecasting models: A critical systematic review. Sustainable Cities and Society 35, 257–270. <https://doi.org/10.1016/j.scs.2017.08.009>
- Laaroussi, Y., Bahrar, M., El Mankibi, M., Draoui, A., Si-Larbi, A., 2020. Occupant presence and behavior: A major issue for building energy performance simulation and assessment. Sustainable Cities and Society 63. <https://doi.org/10.1016/j.scs.2020.102420>
- Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. META group research note 6, 1.
- Lazos, D., Sproul, A.B., Kay, M., 2014. Optimisation of energy management in commercial buildings with weather forecasting inputs: A review. Renewable and Sustainable Energy Reviews 39, 587.
- Lee, E., Rhee, W., 2021. Individualized Short-term Electric Load Forecasting with Deep Neural Network Based Transfer Learning and Meta Learning. IEEE Access. <https://doi.org/10.1109/ACCESS.2021.3053317>
- Lee, S., Kim, Changmin, Park, Y., Son, H., Kim, Changwan, 2011. Data Mining-Based Predictive Model to Determine Project Financial Success Using Project Definition Parameters.
- Lee, W.-S., Lee, K.-P., 2009. Benchmarking the performance of building energy management using data envelopment analysis. Applied Thermal Engineering 29, 3269–3273. <https://doi.org/10.1016/j.applthermaleng.2008.02.034>

- Lei, L., Chen, W., Wu, B., Chen, C., Liu, W., 2021. A building energy consumption prediction model based on rough set theory and deep learning algorithms. *Energy and Buildings* 240, 110886. <https://doi.org/10.1016/j.enbuild.2021.110886>
- Leydesdorff, L., Bornmann, L., Comins, J.A., Milojević, S., 2016. Citations: Indicators of Quality? The Impact Fallacy. *Frontiers in Research Metrics and Analytics* 1.
- Leyh-Bannurah, S.-R., Tian, Z., Karakiewicz, P.I., Wolfgang, U., Sauter, G., Fisch, M., Pehrke, D., Huland, H., Graefen, M., Budäus, L., 2018. Deep Learning for Natural Language Processing in Urology: State-of-the-Art Automated Extraction of Detailed Pathologic Prostate Cancer Data From Narratively Written Electronic Health Records. *JCO Clinical Cancer Informatics* 1–9. <https://doi.org/10.1200/CCI.18.00080>
- Li, C., Ding, Z., Zhao, D., Yi, J., Zhang, G., 2017. Building Energy Consumption Prediction: An Extreme Deep Learning Approach. *Energies* 10, 1525. <https://doi.org/10.3390/en10101525>
- Li, C., Hong, T., Yan, D., 2014. An insight into actual energy use and its drivers in high-performance buildings. *Applied Energy* 131, 394–410. <https://doi.org/10.1016/j.apenergy.2014.06.032>
- Li, C., Tao, Y., Ao, W., Yang, S., Bai, Y., 2018. Improving forecasting accuracy of daily enterprise electricity consumption using a random forest based on ensemble empirical mode decomposition. *Energy* 165, 1220–1227. <https://doi.org/10.1016/j.energy.2018.10.113>
- Li, D., Wang, X., Menassa, C.C., Kamat, V.R., 2020. 12 - Understanding the impact of building thermal environments on occupants' comfort and mental workload demand through human physiological sensing, in: Pacheco-Torgal, F., Rasmussen, E., Granqvist, C.-G., Ivanov, V., Kaklauskas, A., Makonin, S. (Eds.), *Start-Up Creation* (Second Edition), Woodhead Publishing Series in Civil and Structural Engineering. Woodhead Publishing, pp. 291–341. <https://doi.org/10.1016/B978-0-12-819946-6.00012-6>
- Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017. Feature Selection: A Data Perspective. *ACM Comput. Surv.* 50, 94:1-94:45. <https://doi.org/10.1145/3136625>
- Li, K., Xie, X., Xue, W., Dai, X., Chen, X., Yang, X., 2018. A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction. *Energy and Buildings* 174, 323–334. <https://doi.org/10.1016/j.enbuild.2018.06.017>

- 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, 102136. <https://doi.org/10.1016/j.jobr.2020.102136>
- Li, Q., Meng, Q., Cai, J., Yoshino, H., Mochida, A., 2009a. Applying support vector machine to predict hourly cooling load in the building. *Applied Energy* 86, 2249–2256. <https://doi.org/10.1016/j.apenergy.2008.11.035>
- Li, Q., Meng, Q., Cai, J., Yoshino, H., Mochida, A., 2009b. Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. *Energy Conversion and Management* 50, 90–96. <https://doi.org/10.1016/j.enconman.2008.08.033>
- Li, X., Wen, J., 2014. Review of building energy modeling for control and operation. *Renewable and Sustainable Energy Reviews* 37, 517–537. <https://doi.org/10.1016/j.rser.2014.05.056>
- Li, X., Wen, J., Bai, E., 2015. Building energy forecasting using system identification based on system characteristics test. <https://doi.org/10.1109/MSCPES.2015.7115401>
- Li, X., Yao, R., 2020. A machine-learning-based approach to predict residential annual space heating and cooling loads considering occupant behaviour. *Energy* 212, 118676. <https://doi.org/10.1016/j.energy.2020.118676>
- Li, Z., Dong, B., 2017. A new modeling approach for short-term prediction of occupancy in residential buildings. *Building and Environment* 121, 277–290. <https://doi.org/10.1016/j.buildenv.2017.05.005>
- Li, Z.Y., Zhang, S.B., Xiao, Y.B., Shi, Q.Q., Zhao, Y.Q., Gao, J.L., 2018. A new idea of building energy efficiency: The heat transfer coefficient changing with outdoor temperature wall. *IOP Conf. Ser.: Earth Environ. Sci.* 188, 012105. <https://doi.org/10.1088/1755-1315/188/1/012105>
- Liao, J.-M., Chang, M.-J., Chang, L.-M., 2020. Prediction of Air-Conditioning Energy Consumption in R&D Building Using Multiple Machine Learning Techniques. *Energies* 13, 1847. <https://doi.org/10.3390/en13071847>
- Lin, M., Afshari, A., Azar, E., 2018. A data-driven analysis of building energy use with emphasis on operation and maintenance: A case study from the UAE. *Journal of Cleaner Production* 192, 169–178. <https://doi.org/10.1016/j.jclepro.2018.04.270>
- Lin, X., Yu, H., Wang, M., Li, C., Wang, Z., Tang, Y., 2021. Electricity Consumption Forecast of High-Rise Office Buildings Based on the Long Short-Term Memory Method. *Energies* 14, 4785. <https://doi.org/10.3390/en14164785>

- Ling, J., Zhao, L., Xing, J., Lu, Z., 2015. Statistical analysis of residential building energy consumption in Tianjin. *Front. Energy* 8, 513–520. <https://doi.org/10.1007/s11708-014-0327-5>
- Liu, C., Sun, B., Zhang, C., Li, F., 2020. A hybrid prediction model for residential electricity consumption using holt-winters and extreme learning machine. *Applied Energy* 275. <https://doi.org/10.1016/j.apenergy.2020.115383>
- Liu, J., Zhang, Q., Dong, Z., Li, X., Li, G., Xie, Y., Li, K., 2021. Quantitative evaluation of the building energy performance based on short-term energy predictions. *Energy* 223. <https://doi.org/10.1016/j.energy.2021.120065>
- Liu, X., Ding, Y., Tang, H., Fan, L., Lv, J., 2022. Investigating the effects of key drivers on energy consumption of nonresidential buildings: A data-driven approach integrating regularization and quantile regression. *Energy* 244. <https://doi.org/10.1016/j.energy.2021.122720>
- Liu, Y., Chen, H., Zhang, L., Wu, X., Wang, X., 2020a. Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China. *Journal of Cleaner Production* 272, 122542. <https://doi.org/10.1016/j.jclepro.2020.122542>
- Liu, Y., Chen, H., Zhang, L., Wu, X., Wang, X., 2020b. Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China. *Journal of Cleaner Production* 272, 122542. <https://doi.org/10.1016/j.jclepro.2020.122542>
- Loh, W.-Y., 2011. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery* 1, 14–23. <https://doi.org/10.1002/widm.8>
- Lu, S., Wei, S., Zhang, K., Kong, X., Wu, W., 2013. Investigation and analysis on the energy consumption of starred hotel buildings in Hainan Province, the tropical region of China. *Energy Conversion and Management* 75, 570–580. <https://doi.org/10.1016/j.enconman.2013.07.008>
- Luo, J., 2023. Data-Driven Innovation: What Is It? *IEEE Transactions on Engineering Management* 70, 784–790. <https://doi.org/10.1109/TEM.2022.3145231>
- Ma, J.-J., Liu, L.-Q., Su, B., Xie, B.-C., 2015. Exploring the critical factors and appropriate policies for reducing energy consumption of China's urban civil building sector. *Journal of Cleaner Production* 103, 446–454. <https://doi.org/10.1016/j.jclepro.2014.11.001>

- Mafimisebi, I.B., Jones, K., Sennaroglu, B., Nwaubani, S., 2018. A validated low carbon office building intervention model based on structural equation modelling. *Journal of Cleaner Production* 200, 478–489. <https://doi.org/10.1016/j.jclepro.2018.07.249>
- Maggs-Rapport, F., 2001. ‘Best research practice’: in pursuit of methodological rigour. *Journal of Advanced Nursing* 35, 373–383. <https://doi.org/10.1046/j.1365-2648.2001.01853.x>
- Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE* 13, e0194889. <https://doi.org/10.1371/journal.pone.0194889>
- Maldonado, S., Weber, R., 2009. A wrapper method for feature selection using Support Vector Machines. *Information Sciences, Special Section on High Order Fuzzy Sets* 179, 2208–2217. <https://doi.org/10.1016/j.ins.2009.02.014>
- Manfren, M., Nastasi, B., Tronchin, L., 2020. Linking design and operation phase energy performance analysis through regression-based approaches. *Frontiers in Energy Research* 8. <https://doi.org/10.3389/fenrg.2020.557649>
- Mangula, M., 2019. Modeling Sustainability of Energy Access in Rural Areas of Tanzania.
- Marino, C., Nucara, A., Pietrafesa, M., 2017. Does window-to-wall ratio have a significant effect on the energy consumption of buildings? A parametric analysis in Italian climate conditions. *Journal of Building Engineering* 13, 169–183. <https://doi.org/10.1016/j.jobe.2017.08.001>
- Marrow, J. (2023) *Guide to Controls (BG83/2023)*. Available at: https://www.bsria.com/uk/product/DJLLPr/guide_to_controls_bg832023_a15d25e1/?_gl=1*tL27fg*_up*MQ..*_ga*OTUxODE1ODYyLjE3MzEwMTE0NTE.*_ga_L0PN3DGN1B*MTczMTAxMTQ1MC4xLjAuMTczMTAxMTQ1MC4wLjAuMA.. (Accessed: 7 November 2024).
- Martin, R.F., 2000. General Deming Regression for Estimating Systematic Bias and Its Confidence Interval in Method-Comparison Studies. *Clinical Chemistry* 46, 100–104. <https://doi.org/10.1093/clinchem/46.1.100>
- Marwan, M., 2020. The effect of wall material on energy cost reduction in building. *Case Studies in Thermal Engineering* 17, 100573. <https://doi.org/10.1016/j.csite.2019.100573>
- Mat Daut, M.A., Hassan, M.Y., Abdullah, H., Rahman, H.A., Abdullah, M.P., Hussin, F., 2017. Building electrical energy consumption forecasting analysis using conventional

- and artificial intelligence methods: A review. Renewable and Sustainable Energy Reviews 70, 1108–1118. <https://doi.org/10.1016/j.rser.2016.12.015>
- Matveeva, O., Nechipurenko, Y., Rossi, L., Moore, B., Sætrom, P., Ogurtsov, A.Y., Atkins, J.F., Shabalina, S.A., 2007. Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. Nucleic Acids Research 35, e63. <https://doi.org/10.1093/nar/gkm088>
- Mavromatidis, L.E., Marsault, X., Lequay, H., 2014. Daylight factor estimation at an early design stage to reduce buildings' energy consumption due to artificial lighting: A numerical approach based on Doehlert and Box-Behnken designs. Energy 65, 488–502. <https://doi.org/10.1016/j.energy.2013.12.028>
- Mawson, V.J., Hughes, B.R., 2020. Deep learning techniques for energy forecasting and condition monitoring in the manufacturing sector. Energy and Buildings 217, 109966. <https://doi.org/10.1016/j.enbuild.2020.109966>
- Mazzeo, D., Kontoleon, K.J., 2020. The role of inclination and orientation of different building roof typologies on indoor and outdoor environment thermal comfort in Italy and Greece. Sustainable Cities and Society 60. <https://doi.org/10.1016/j.scs.2020.102111>
- Melo, A.P., Cóstola, D., Lamberts, R., Hensen, J.L.M., 2014. Development of surrogate models using artificial neural network for building shell energy labelling. Energy Policy 69, 457–466. <https://doi.org/10.1016/j.enpol.2014.02.001>
- Meng, Q., Xiong, C., Mourshed, M., Wu, M., Ren, X., Wang, W., Li, Y., Song, H., 2020. Change-point multivariable quantile regression to explore effect of weather variables on building energy consumption and estimate base temperature range. Sustainable Cities and Society 53. <https://doi.org/10.1016/j.scs.2019.101900>
- Milivojevic, N. and Ahmed, A. (2018) *Evaluating learning management mechanisms and requirements for achieving BIM competencies: an in-depth study of ACE practitioners.*
- Miller, Z., Dickinson, B., Hu, W., 2012. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features 2012. <https://doi.org/10.4236/ijis.2012.224019>
- Min, J., Azevedo, I.L., Hakkarainen, P., 2015. Assessing regional differences in lighting heat replacement effects in residential buildings across the United States. Applied Energy 141, 12–18. <https://doi.org/10.1016/j.apenergy.2014.11.031>

- Mishra, P., Biancolillo, A., Roger, J.M., Marini, F., Rutledge, D.N., 2020. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends in Analytical Chemistry* 132, 116045.
<https://doi.org/10.1016/j.trac.2020.116045>
- Mocanu, E., Nguyen, P.H., Gibescu, M., Kling, W.L., 2016. Deep learning for estimating building energy consumption. *Sustainable Energy, Grids and Networks* 6, 91–99.
<https://doi.org/10.1016/j.segan.2016.02.005>
- Moghaddam, A.H., Moghaddam, M.H., Esfandyari, M., 2016. Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science* 21, 89–93. <https://doi.org/10.1016/j.jefas.2016.07.002>
- Moghram, I., Rahman, S., 1989. Analysis and evaluation of five short-term load forecasting techniques. *IEEE Transactions on Power Systems; (USA)* 4:4.
<https://doi.org/10.1109/59.41700>
- Mohanty, D., Palai, A.K., 2023. Comprehensive Machine Learning Pipeline for Prediction of Power Conversion Efficiency in Perovskite Solar Cells. *Advanced Theory and Simulations* 6, 2300309. <https://doi.org/10.1002/adts.202300309>
- Molina-Solana, M., Ros, M., Ruiz, M.D., Gómez-Romero, J., Martin-Bautista, M.J., 2017. Data science for building energy management: A review. *Renewable and Sustainable Energy Reviews* 70, 598–609. <https://doi.org/10.1016/j.rser.2016.11.132>
- Nasteski, V., 2017. An overview of the supervised machine learning methods. *HORIZONS.B* 4, 51–62. <https://doi.org/10.20544/HORIZONS.B.04.1.17.P05>
- Neto, A.H., Fiorelli, F.A.S., 2008. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings* 40, 2169–2176. <https://doi.org/10.1016/j.enbuild.2008.06.013>
- Newgard, C.D., Lewis, R.J., 2015. Missing Data: How to Best Account for What Is Not Known. *JAMA* 314, 940–941. <https://doi.org/10.1001/jama.2015.10516>
- Newman, D.R., Cockburn, J.M.H., Drăguț, L., Lindsay, J.B., 2022. Local scale optimization of geomorphometric land surface parameters using scale-standardized Gaussian scale-space. *Computers & Geosciences* 165, 105144.
<https://doi.org/10.1016/j.cageo.2022.105144>
- Newman, I., Benz, C.R., Ridenour, C.S., 1998. Qualitative-quantitative Research Methodology: Exploring the Interactive Continuum. SIU Press.

- Ngarambe, J., Irakoze, A., Yun, G.Y., Kim, G., 2020. Comparative performance of machine learning algorithms in the prediction of indoor daylight illuminances. *Sustainability* (Switzerland) 12. <https://doi.org/10.3390/su12114471>
- Nguyen, C., Wang, Y., Nguyen, H.N., 2013. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic 2013. <https://doi.org/10.4236/jbise.2013.65070>
- Niu, D., Wang, Y., Wu, D.D., 2010. Power load forecasting using support vector machine and ant colony optimization. *Expert Systems with Applications* 37, 2531–2539. <https://doi.org/10.1016/j.eswa.2009.08.019>
- Noh, B., Son, J., Park, H., Chang, S., 2017. In-depth analysis of energy efficiency related factors in commercial buildings using data cube and association rule mining. *Sustainability* (Switzerland) 9. <https://doi.org/10.3390/su9112119>
- Ocampo Batlle, E.A., Escobar Palacio, J.C., Silva Lora, E.E., Martínez Reyes, A.M., Melian Moreno, M., Morejón, M.B., 2020. A methodology to estimate baseline energy use and quantify savings in electrical energy consumption in higher education institution buildings: Case study, Federal University of Itajubá (UNIFEI). *Journal of Cleaner Production* 244. <https://doi.org/10.1016/j.jclepro.2019.118551>
- Office for National Statistics, 2019. Overview of the UK population - Office for National Statistics [WWW Document]. URL <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/january2021> (accessed 4.8.21).
- Olawumi, T.O., Chan, D.W.M., Ojo, S., Yam, M.C.H., 2022. Automating the modular construction process: A review of digital technologies and future directions with blockchain technology. *Journal of Building Engineering* 46, 103720. <https://doi.org/10.1016/j.jobe.2021.103720>
- Olu-Ajayi, R., 2017. An Investigation into the Suitability of k-Nearest Neighbour (k-NN) for Software Effort Estimation. *ijacsa* 8. <https://doi.org/10.14569/IJACSA.2017.080628>
- Olu-Ajayi, R., Alaka, H., 2021. Building energy consumption prediction using deep learning. Environmental Design and Management Conference (EDMIC).
- Olu-Ajayi, R., Alaka, H., Owolabi, H., Akanbi, L., Ganiyu, S., 2023a. Data-Driven Tools for Building Energy Consumption Prediction: A Review. *Energies* 16, 2574. <https://doi.org/10.3390/en16062574>
- Olu-Ajayi, R., Alaka, H., Sulaimon, I., Balogun, H., Wusu, G., Yusuf, W., Adegoke, M., 2023b. Building energy performance prediction: A reliability analysis and evaluation

- of feature selection methods. *Expert Systems with Applications* 225, 120109.
<https://doi.org/10.1016/j.eswa.2023.120109>
- Olu-Ajayi, R., Alaka, H., Sulaimon, I., Grishikashvili, K., Sunmola, F., Oseghale, R., Ajayi, S., 2021. Ensemble learning for energy performance prediction of residential buildings. Environmental Design and Management Conference (EDMIC).
- Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., Ajayi, S., 2022a. Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering* 45, 103406.
<https://doi.org/10.1016/j.jobe.2021.103406>
- Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., Ajayi, S., 2022b. Machine learning for energy performance prediction at the design stage of buildings. *Energy for Sustainable Development* 66, 12–25. <https://doi.org/10.1016/j.esd.2021.11.002>
- Ortiz-Bejar, José, Graff, M., Tellez, E.S., Ortiz-Bejar, Jesús, Jacobo, J.C., 2018. k-Nearest Neighbor Regressors Optimized by using Random Search, in: 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). Presented at the 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), pp. 1–5. <https://doi.org/10.1109/ROPEC.2018.8661399>
- Ozaki, Y., Yano, M., Onishi, M., 2017. Effective hyperparameter optimization using Nelder-Mead method in deep learning. *IPSJ T Comput Vis Appl* 9, 20.
<https://doi.org/10.1186/s41074-017-0030-7>
- Pan, Y., Zhang, L., 2020. Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Applied Energy* 268.
<https://doi.org/10.1016/j.apenergy.2020.114965>
- Pang, Z., O'Neill, Z., 2018. Uncertainty quantification and sensitivity analysis of the domestic hot water usage in hotels. *Applied Energy* 232, 424–442.
<https://doi.org/10.1016/j.apenergy.2018.09.221>
- Paone, A., Bacher, J.-P., 2018. The Impact of Building Occupant Behavior on Energy Efficiency and Methods to Influence It: A Review of the State of the Art. *Energies* 11, 953. <https://doi.org/10.3390/en11040953>
- Parhizkar, T., Rafieipour, E., Parhizkar, A., 2021. Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction. *Journal of Cleaner Production* 279, 123866.
<https://doi.org/10.1016/j.jclepro.2020.123866>

- Park, J.H., Wi, S., Chang, S.J., Kim, S., 2020. Analysis of energy retrofit system using latent heat storage materials applied to residential buildings considering climate impacts. *Applied Thermal Engineering* 169.
<https://doi.org/10.1016/j.applthermaleng.2020.114904>
- Paudel, S., Elmitri, M., Couturier, S., Nguyen, P.H., Kamphuis, R., Lacarrière, B., Le Corre, O., 2017. A relevant data selection method for energy consumption prediction of low energy building based on support vector machine. *Energy and Buildings* 138, 240–256. <https://doi.org/10.1016/j.enbuild.2016.11.009>
- Paukštys, V., Cinelis, G., Mockienė, J., Daukšys, M., 2021. Airtightness and heat energy loss of mid-size terraced houses built of different construction materials. *Energies* 14.
<https://doi.org/10.3390/en14196367>
- Pavlenko, D.V. (2019) ‘REVIEW OF THE OPTIMIZATION THEORY’, Студентство. Наука. Іноземна мова: збірник наукових праць, p. 79.
- Pawar, A., Jaiswal, D.R.C., 2020. STOCK MARKET STUDY USING SUPERVISED MACHINE LEARNING 4.
- Peng, B., Zou, H.-M., Bai, P.-F., Feng, Y.-Y., 2021. Building energy consumption prediction and energy control of large-scale shopping malls based on a noncentralized self-adaptive energy management control system. *Energy Exploration and Exploitation* 39, 1381–1393. <https://doi.org/10.1177/0144598720920731>
- Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., Ben Taieb, S., Bergmeir, C., Bessa, R.J., Bijak, J., Boylan, J.E., Browell, J., Carnevale, C., Castle, J.L., Cirillo, P., Clements, M.P., Cordeiro, C., Cyrino Oliveira, F.L., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P.H., Frazier, D.T., Gilliland, M., Gönül, M.S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, Mariangela, Guidolin, Massimo, Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D.F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V.R.R., Kang, Y., Koehler, A.B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G.M., Martinez, A.B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J.M., Pedregal, D.J., Pinson, P., Ramos, P., Rapach, D.E., Reade, J.J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H.L., Spiliotis, E., Syntetos, A.A., Talagala, P.D., Talagala, T.S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Trapero Arenas, J.R., Wang, X., Winkler, R.L., Yusupova, A., Ziel, F., 2022. Forecasting: theory and

- practice. *International Journal of Forecasting* 38, 705–871.
<https://doi.org/10.1016/j.ijforecast.2021.11.001>
- 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>
- Phillips, V., Barker, E., 2021. Systematic reviews: Structure, form and content. *Journal of Perioperative Practice* 31, 349–353. <https://doi.org/10.1177/1750458921994693>
- Pinheiro, G., Pereira, T., Dias, C., Freitas, C., Hespanhol, V., Costa, J.L., Cunha, A., Oliveira, H.P., 2020. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Sci Rep* 10, 3625.
<https://doi.org/10.1038/s41598-020-60202-3>
- Pino-Mejías, R., Pérez-Fargallo, A., Rubio-Bellido, C., Pulido-Arcas, J.A., 2017. Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO₂ emissions. *Energy* 118, 24–36.
<https://doi.org/10.1016/j.energy.2016.12.022>
- Pirbazari, A.M., Chakravorty, A., Rong, C., 2019. Evaluating Feature Selection Methods for Short-Term Load Forecasting, in: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). Presented at the 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 1–8.
<https://doi.org/10.1109/BIGCOMP.2019.8679188>
- Pisello, A.L., Castaldo, V.L., Piselli, C., Fabiani, C., Cotana, F., 2017. Thermal performance of coupled cool roof and cool façade: Experimental monitoring and analytical optimization procedure. *Energy and Buildings* 157, 35–52.
<https://doi.org/10.1016/j.enbuild.2017.04.054>
- Ponta, L., Puliga, G., Oneto, L., Manzini, R., 2022. Identifying the Determinants of Innovation Capability With Machine Learning and Patents. *IEEE Transactions on Engineering Management* 69, 2144–2154.
<https://doi.org/10.1109/TEM.2020.3004237>
- Pora, U., Gerdtsri, N., Thawesaengskulthai, N., Triukose, S., 2022. Data-Driven Roadmapping (DDRM): Approach and Case Demonstration. *IEEE Transactions on Engineering Management* 69, 209–227. <https://doi.org/10.1109/TEM.2020.3005341>

- Premrov, M., Žigart, M., Žegarac Leskovar, V., 2018. Influence of the building shape on the energy performance of timber-glass buildings located in warm climatic regions. *Energy* 149, 496–504. <https://doi.org/10.1016/j.energy.2018.02.074>
- Public Health England, 2019. Public Health England publishes air pollution evidence review [WWW Document]. GOV.UK. URL <https://www.gov.uk/government/news/public-health-england-publishes-air-pollution-evidence-review> (accessed 4.8.21).
- Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W., O’Sullivan, J.M., 2022. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics* 2.
- Qiao, Q., Yunusa-Kaltungo, A., Edwards, R., 2020a. Hybrid method for building energy consumption prediction based on limited data, in: 2020 IEEE PES/IAS PowerAfrica. Presented at the 2020 IEEE PES/IAS PowerAfrica, pp. 1–5. <https://doi.org/10.1109/PowerAfrica49420.2020.9219915>
- Qiao, Q., Yunusa-Kaltungo, A., Edwards, R., 2020b. Predicting building energy consumption based on meteorological data, in: 2020 IEEE PES/IAS PowerAfrica. Presented at the 2020 IEEE PES/IAS PowerAfrica, pp. 1–5. <https://doi.org/10.1109/PowerAfrica49420.2020.9219909>
- Qiao, Q., Yunusa-Kaltungo, A., Edwards, R.E., 2021. Towards developing a systematic knowledge trend for building energy consumption prediction. *Journal of Building Engineering* 35, 101967. <https://doi.org/10.1016/j.jobe.2020.101967>
- Qiong Li, Peng Ren, Qinglin Meng, 2010. Prediction model of annual energy consumption of residential buildings, in: 2010 International Conference on Advances in Energy Engineering. Presented at the 2010 International Conference on Advances in Energy Engineering, pp. 223–226. <https://doi.org/10.1109/ICAEE.2010.5557576>
- Qiu, C., Yi, Y.K., Wang, M., Yang, H., 2020. Coupling an artificial neuron network daylighting model and building energy simulation for vacuum photovoltaic glazing. *Applied Energy* 263. <https://doi.org/10.1016/j.apenergy.2020.114624>
- Rahul, K., Seth, N., Kumar, U.D., 2018. Spotting earnings manipulation: using machine learning for financial fraud detection, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer, pp. 343–356.
- Ravi, A., 2020. Stacked Generalization for Human Activity Recognition. arXiv:2009.10312 [cs].

- Raza, M.Q., Khosravi, A., 2015. A review on artificial 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>
- Reichert, J., Arnold, A.L., Hoogenboom, M.O., Schubert, P., Wilke, T., 2019. Impacts of microplastics on growth and health of hermatypic corals are species-specific. *Environmental Pollution* 254, 113074. <https://doi.org/10.1016/j.envpol.2019.113074>
- Ren, J., Zhou, X., An, J., Yan, D., Shi, X., Jin, X., Zheng, S., 2021. Comparative analysis of window operating behavior in three different open-plan offices in Nanjing. *Energy and Built Environment* 2, 175–187. <https://doi.org/10.1016/j.enbenv.2020.07.007>
- Ríos Canales, V., 2016. Using a Supervised Learning Model: Two-Class Boosted Decision Tree Algorithm for Income Prediction.
- Robinson, Caleb, Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M.A., Pendyala, R.M., 2017a. Machine learning approaches for estimating commercial building energy consumption. *Applied Energy* 208, 889–904. <https://doi.org/10.1016/j.apenergy.2017.09.060>
- Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M.A., Pendyala, R.M., 2017. Machine learning approaches for estimating commercial building energy consumption. *Applied Energy* 208, 889–904. <https://doi.org/10.1016/j.apenergy.2017.09.060>
- Robinson, Caleb, Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M.A., Pendyala, R.M., 2017b. Machine learning approaches for estimating commercial building energy consumption. *Applied Energy* 208, 889–904. <https://doi.org/10.1016/j.apenergy.2017.09.060>
- Rouleau, J., Gosselin, L., Blanchet, P., 2018. Understanding energy consumption in high-performance social housing buildings: A case study from Canada. *Energy* 145, 677–690. <https://doi.org/10.1016/j.energy.2017.12.107>
- Runge, J., Zmeureanu, R., 2019. Forecasting Energy Use in Buildings Using Artificial Neural Networks: A Review. *Energies* 12, 3254. <https://doi.org/10.3390/en12173254>
- Rupf, R., Tsai, M.C., Thomas, S.G., Klimstra, M., 2021. Original article: Validity of measuring wheelchair kinematics using one inertial measurement unit during commonly used testing protocols in elite wheelchair court sports. *Journal of Biomechanics* 127, 110654. <https://doi.org/10.1016/j.jbiomech.2021.110654>
- Rusek, R., Melendez Frigola, J., Colomer Llinas, J., 2022. Influence of occupant presence patterns on energy consumption and its relation to comfort: a case study based on

- sensor and crowd-sensed data. *Energy, Sustainability and Society* 12.
<https://doi.org/10.1186/s13705-022-00336-6>
- Russell, S., Norvig, P., 2020. *Artificial Intelligence: A Modern Approach*, 4th Edition.
- Sadeghi, A., Sinaki, R.Y., Young, W.A., II, Weckman, G.R., 2020. An intelligent model to predict energy performances of residential buildings based on deep neural networks. *Energies* 13. <https://doi.org/10.3390/en13030571>
- Sadineni, S.B., Madala, S., Boehm, R.F., 2011. Passive building energy savings: A review of building envelope components. *Renewable and Sustainable Energy Reviews* 15, 3617–3631. <https://doi.org/10.1016/j.rser.2011.07.014>
- Safa, M., Safa, M., Allen, J., Shahi, A., Haas, C.T., 2017. Improving sustainable office building operation by using historical data and linear models to predict energy usage. *Sustainable Cities and Society* 29, 107–117. <https://doi.org/10.1016/j.scs.2016.12.001>
- Saka, A.B., Chan, D.W.M., 2019. A Scientometric Review and Metasynthesis of Building Information Modelling (BIM) Research in Africa. *Buildings* 9, 85.
<https://doi.org/10.3390/buildings9040085>
- Santos, M.S., Soares, J.P., Abreu, P.H., Araujo, H., Santos, J., 2018. Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. *IEEE Computational Intelligence Magazine* 13, 59–76.
<https://doi.org/10.1109/MCI.2018.2866730>
- Saunders, M., Lewis, P., Thornhill, A., 2009. *Research Methods for Business Students*. Pearson Education.
- Schlender, J.-F., Teutonico, D., Coboeken, K., Schnizler, K., Eissing, T., Willmann, S., Jaehde, U., Stass, H., 2018. A Physiologically-Based Pharmacokinetic Model to Describe Ciprofloxacin Pharmacokinetics Over the Entire Span of Life. *Clin Pharmacokinet* 57, 1613–1634. <https://doi.org/10.1007/s40262-018-0661-6>
- Schlosser, R.W., 2007. Appraising the quality of systematic reviews. *Focus* 17, 1–8.
- Seijo-Pardo, B., Bolón-Canedo, V., Alonso-Betanzos, A., 2019. On developing an automatic threshold applied to feature selection ensembles. *Information Fusion* 45, 227–245.
<https://doi.org/10.1016/j.inffus.2018.02.007>
- Serale, G., Fiorentini, M., Noussan, M., 2020. 11 - Development of algorithms for building energy efficiency, in: Pacheco-Torgal, F., Rasmussen, E., Granqvist, C.-G., Ivanov, V., Kaklauskas, A., Makonin, S. (Eds.), *Start-Up Creation (Second Edition)*, Woodhead Publishing Series in Civil and Structural Engineering. Woodhead Publishing, pp. 267–290. <https://doi.org/10.1016/B978-0-12-819946-6.00011-4>

- 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, 5. <https://doi.org/10.1186/s40327-018-0064-7>
- Seyedzadeh, S., Rahimian, F.P., Oliver, S., Glesk, I., Kumar, B., 2020. Data driven model improved by multi-objective optimisation for prediction of building energy loads. *Automation in Construction* 116, 103188. <https://doi.org/10.1016/j.autcon.2020.103188>
- Sha, H., Xu, P., Hu, C., Li, Z., Chen, Y., Chen, Z., 2019. A simplified HVAC energy prediction method based on degree-day. *Sustainable Cities and Society* 51. <https://doi.org/10.1016/j.scs.2019.101698>
- Shakya, S., Choosong, T., Techato, K., Gyawali, S., Panthee, B., Shrestha, N., Dangal, M.R., 2021. Indoor Air Pollution (IAP) Traceable to Household Fuel Consumption and Its Impact on Health. *Kathmandu Univ Med J* 73, 123–31.
- Shan, S., Cao, B., Wu, Z., 2019. Forecasting the Short-Term Electricity Consumption of Building Using a Novel Ensemble Model. *IEEE Access* 7, 88093–88106. <https://doi.org/10.1109/ACCESS.2019.2925740>
- Shao, M., Wang, X., Bu, Z., Chen, X., Wang, Y., 2020. Prediction of energy consumption in hotel buildings via support vector machines. *Sustainable Cities and Society* 57, 102128. <https://doi.org/10.1016/j.scs.2020.102128>
- Shao, X., Pu, C., Zhang, Y., Kim, C.S., 2020. Domain Fusion CNN-LSTM for Short-Term Power Consumption Forecasting. *IEEE Access* 8, 188352–188362. <https://doi.org/10.1109/ACCESS.2020.3031958>
- Shapi, M.K.M., Ramli, N.A., Awalin, L.J., 2021. Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment* 5, 100037. <https://doi.org/10.1016/j.dibe.2020.100037>
- Sharma, J., Giri, C., Granmo, O.-C., Goodwin, M., 2019. Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. *EURASIP J. on Info. Security* 2019, 15. <https://doi.org/10.1186/s13635-019-0098-y>
- Shen, M., Lu, Y., Wei, K.H., Cui, Q., 2020. Prediction of household electricity consumption and effectiveness of concerted intervention strategies based on occupant behaviour and personality traits. *Renewable and Sustainable Energy Reviews* 127, 109839. <https://doi.org/10.1016/j.rser.2020.109839>

- Shen, Y., Liu, D., Jiang, L., Nielsen, K., Yin, J., Liu, J., Bauer-Gottwein, P., 2022. High-resolution water level and storage variation datasets for 338 reservoirs in China during 2010–2021. *Earth System Science Data* 14, 5671–5694. <https://doi.org/10.5194/essd-14-5671-2022>
- Silvestro, F., Bagnasco, A., Lanza, I., Massucco, S., Vinci, A., 2017. Energy efficient policy and real time energy monitoring in a large hospital facility: A case study. *International Journal of Heat and Technology* 35, S221–S227.
<https://doi.org/10.18280/ijht.35Sp0131>
- Singh, M.M., Singaravel, S., Geyer, P., 2021. Machine learning for early stage building energy prediction: Increment and enrichment. *Applied Energy* 304, 117787.
<https://doi.org/10.1016/j.apenergy.2021.117787>
- Skeie, K., Gustavsen, A., 2021. Utilising open geospatial data to refine weather variables for building energy performance evaluation—incident solar radiation and wind-driven infiltration modelling. *Energies* 14. <https://doi.org/10.3390/en14040802>
- Smith, J.S., Nebgen, B.T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., Roitberg, A.E., 2019. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* 10, 2903. <https://doi.org/10.1038/s41467-019-10827-4>
- Socolow, R.H., 1978. The twin rivers program on energy conservation in housing: Highlights and conclusions. *Energy and Buildings* 1, 207–242. [https://doi.org/10.1016/0378-7788\(78\)90003-8](https://doi.org/10.1016/0378-7788(78)90003-8)
- Somu, N., M R, G.R., Ramamritham, K., 2020. A hybrid model for building energy consumption forecasting using long short term memory networks. *Applied Energy* 261. <https://doi.org/10.1016/j.apenergy.2019.114131>
- Somu, N., Raman M R, G., Ramamritham, K., 2021. A deep learning framework for building energy consumption forecast. *Renewable and Sustainable Energy Reviews* 137.
<https://doi.org/10.1016/j.rser.2020.110591>
- Sonkamble, B.A., Doye, D.D., 2008. An overview of speech recognition system based on the support vector machines, in: 2008 International Conference on Computer and Communication Engineering. Presented at the 2008 International Conference on Computer and Communication Engineering, pp. 768–771.
<https://doi.org/10.1109/ICCCE.2008.4580709>

- Spiliotis, E., Kouloumos, A., Assimakopoulos, V., Makridakis, S., 2020. Are forecasting competitions data representative of the reality? International Journal of Forecasting, M4 Competition 36, 37–53. <https://doi.org/10.1016/j.ijforecast.2018.12.007>
- Srivastava, S., Gupta, M.R., Frigyik, B.A., 2007. Bayesian quadratic discriminant analysis. Journal of Machine Learning Research 8.
- Su, X., Zhang, L., Luo, Y., Liu, Z., Yang, H., Wang, X., 2021. Conceptualization and preliminary analysis of a novel reversible photovoltaic window. Energy Conversion and Management 250. <https://doi.org/10.1016/j.enconman.2021.114925>
- Suh, D., Chang, S., 2014. A heuristic rule-based passive design decision model for reducing heating energy consumption of Korean apartment buildings. Energies 7, 6897–6929. <https://doi.org/10.3390/en7116897>
- Sulaimon, I., Alaka, H., Olu-Ajayi, R., Ahmad, M., Sunmola, F., Ajayi, S., Hye, A., 2021. Air Pollution Prediction using Machine Learning – A Review.
- Sumaiya Thaseen, I., Aswani Kumar, C., 2017. Intrusion detection model using fusion of chi-square feature selection and multi class SVM. Journal of King Saud University - Computer and Information Sciences 29, 462–472. <https://doi.org/10.1016/j.jksuci.2015.12.004>
- Swanson, N.R., Xiong, W., 2018. Big data analytics in economics: What have we learned so far, and where should we go from here? Canadian Journal of Economics/Revue canadienne d'économique 51, 695–746. <https://doi.org/10.1111/caje.12336>
- Szul, T., Tabor, S., Pancerz, K., 2021. Application of the BORUTA algorithm to input data selection for a model based on rough set theory (RST) to prediction energy consumption for building heating. Energies 14. <https://doi.org/10.3390/en14102779>
- Tafakkori, R., Fattahi, A., 2021. Introducing novel configurations for double-glazed windows with lower energy loss. Sustainable Energy Technologies and Assessments 43. <https://doi.org/10.1016/j.seta.2020.100919>
- Tahmasebi, M.M., Banihashemi, S., Hassanabadi, M.S., 2011. Assessment of the Variation Impacts of Window on Energy Consumption and Carbon Footprint. Procedia Engineering, 2011 International Conference on Green Buildings and Sustainable Cities 21, 820–828. <https://doi.org/10.1016/j.proeng.2011.11.2083>
- Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., Finn, D., 2015. Data Driven Approaches for Prediction of Building Energy Consumption at Urban Level. Energy Procedia 78, 3378–3383. <https://doi.org/10.1016/j.egypro.2015.11.754>
- The World Bank, 2019. Population, total | Data.

- TOPRAK, Ahmet, KOKLU, N., TOPRAK, Aysegul, OZCAN, R., 2017. International Journal of Intelligent Systems and Applications in Engineering.
- Tranfield, D., Denyer, D., Smart, P., 2003. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management* 14, 207–222. <https://doi.org/10.1111/1467-8551.00375>
- Truong, N.-S., Ngo, N.-T., Pham, A.-D., 2021. Forecasting Time-Series Energy Data in Buildings Using an Additive Artificial Intelligence Model for Improving Energy Efficiency. *Comput Intell Neurosci* 2021, 6028573. <https://doi.org/10.1155/2021/6028573>
- Tsai, C.-F., Hsu, Y.-F., Yen, D.C., 2014. A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing* 24, 977–984. <https://doi.org/10.1016/j.asoc.2014.08.047>
- Tso, G.K.F., Yau, K.K.W., 2007. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* 32, 1761–1768. <https://doi.org/10.1016/j.energy.2006.11.010>
- Tussen, R.J.W., Buter, R.K., van Leeuwen, Th.N., 2000. Technological Relevance of Science: An Assessment of Citation Linkages between Patents and Research Papers. *Scientometrics* 47, 389–412. <https://doi.org/10.1023/A:1005603513439>
- United Nations Environment Programme, U.N., 2017. Sustainable buildings [WWW Document]. UNEP - UN Environment Programme. URL <http://www.unep.org/explore-topics/resource-efficiency/what-we-do/cities/sustainable-buildings> (accessed 3.16.21).
- United Nations Framework Convention on Climate Change, 2015. The Paris Agreement | UNFCCC [WWW Document]. URL <https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement> (accessed 3.16.21).
- United Nations Population Division, 2017. World Population Prospects - Population Division - United Nations [WWW Document]. URL <https://population.un.org/wpp/> (accessed 3.4.22).
- Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical Combinations and Scientific Impact. *Science* 342, 468–472. <https://doi.org/10.1126/science.1240474>
- Van Eck, N.J., Waltman, L., 2020. VOSviewer Manual: Manual for VOSviewer version 1.6. 15. Leiden: Centre for Science and Technology Studies (CWTS) of Leiden University.

- van Eck, N.J., Waltman, L., 2014. Visualizing Bibliometric Networks, in: Ding, Y., Rousseau, R., Wolfram, D. (Eds.), *Measuring Scholarly Impact: Methods and Practice*. Springer International Publishing, Cham, pp. 285–320.
https://doi.org/10.1007/978-3-319-10377-8_13
- Verma, A., Prakash, S., Kumar, A., Aghamohammadi, N., 2022. A novel design approach for indoor environmental quality based on a multiagent system for intelligent buildings in a smart city: Toward occupant's comfort. *Environmental Progress and Sustainable Energy*. <https://doi.org/10.1002/ep.13895>
- Verma, S.K., Anand, Y., Gupta, N., Jindal, B.B., Tyagi, V.V., Anand, S., 2022. Hygrothermal dynamics for developing energy-efficient buildings: Building materials and ventilation system considerations. *Energy and Buildings* 260, 111932.
<https://doi.org/10.1016/j.enbuild.2022.111932>
- Veugelers, R., Wang, J., 2019. Scientific novelty and technological impact. *Research Policy* 48, 1362–1372. <https://doi.org/10.1016/j.respol.2019.01.019>
- Vorobeychik, Y., Wallrabenstein, J.R., 2013. Using Machine Learning for Operational Decisions in Adversarial Environments 9.
- Wang, E., 2017. Decomposing core energy factor structure of U.S. commercial buildings through clustering around latent variables with Random Forest on large-scale mixed data. *Energy Conversion and Management* 153, 346–361.
<https://doi.org/10.1016/j.enconman.2017.10.020>
- Wang, J., Chen, X., Zhang, F., Chen, F., Xin, Y., 2021. Building Load Forecasting Using Deep Neural Network with Efficient Feature Fusion. *Journal of Modern Power Systems and Clean Energy* 9, 160–169. <https://doi.org/10.35833/MPCE.2020.000321>
- Wang, Jinsong, Chen, X., Zhang, F., Chen, F., Xin, Y., 2021. Building Load Forecasting Using Deep Neural Network with Efficient Feature Fusion. *Journal of Modern Power Systems and Clean Energy* 9, 160–169. <https://doi.org/10.35833/MPCE.2020.000321>
- Wang, J., Li, Z., Tam, V.W.Y., 2014. Critical factors in effective construction waste minimization at the design stage: A Shenzhen case study, China. *Resources, Conservation and Recycling* 82, 1–7. <https://doi.org/10.1016/j.resconrec.2013.11.003>
- Wang, J., Plataniotis, K.N., Lu, J., Venetsanopoulos, A.N., 2008. Kernel quadratic discriminant analysis for small sample size problem. *Pattern Recognition* 41, 1528–1538. <https://doi.org/10.1016/j.patcog.2007.10.024>

- Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy* 46, 1416–1436.
<https://doi.org/10.1016/j.respol.2017.06.006>
- Wang, N., 2017. Bankruptcy Prediction Using Machine Learning. *Journal of Mathematical Finance* 07, 908. <https://doi.org/10.4236/jmf.2017.74049>
- Wang, R., Cao, Q., Zhao, Q., Li, Y., 2018. Bioindustry in China: An overview and perspective. *New Biotechnology, Bioeconomy* 40, 46–51.
<https://doi.org/10.1016/j.nbt.2017.08.002>
- Wang, R., Lu, S., Feng, W., 2020. A novel improved model for building energy consumption prediction based on model integration. *Applied Energy* 262, 114561.
<https://doi.org/10.1016/j.apenergy.2020.114561>
- Wang, R., Lu, S., Li, Q., 2019. Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustainable Cities and Society* 49, 101623. <https://doi.org/10.1016/j.scs.2019.101623>
- Wang, R., Zhao, H., Wu, Y., Wang, Y., Feng, X., Liu, M., 2018. An industrial facility layout design method considering energy saving based on surplus rectangle fill algorithm. *Energy* 158, 1038–1051. <https://doi.org/10.1016/j.energy.2018.06.105>
- Wang, W., Zmeureanu, R., Rivard, H., 2005. Applying multi-objective genetic algorithms in green building design optimization. *Building and Environment* 40, 1512–1525.
<https://doi.org/10.1016/j.buildenv.2004.11.017>
- Wang, Z., Srinivasan, R.S., 2017. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews* 75, 796–808.
<https://doi.org/10.1016/j.rser.2016.10.079>
- Wang, Z., Wang, Y., Srinivasan, R.S., 2018a. A novel ensemble learning approach to support building energy use prediction. *Energy and Buildings* 159, 109–122.
<https://doi.org/10.1016/j.enbuild.2017.10.085>
- Wang, Z., Wang, Y., Zeng, R., Srinivasan, R.S., Ahrentzen, S., 2018b. Random Forest based hourly building energy prediction. *Energy and Buildings* 171, 11–25.
<https://doi.org/10.1016/j.enbuild.2018.04.008>
- Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., Zhao, X., 2018. A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews* 82, 1027–1047.
<https://doi.org/10.1016/j.rser.2017.09.108>

- Weingart, S., Guo, H., Börner, K., 2010. Science of Science (Sci 2) Tool User Manual, Version Alpha 3.
- William, M.A., Suárez-López, M.J., Soutullo, S., Hanafy, A.A., 2021. Techno-economic evaluation of building envelope solutions in hot arid climate: A case study of educational building. *Energy Reports* 7, 550–558.
<https://doi.org/10.1016/j.egyr.2021.07.098>
- Wood, D.A., 2022. Chapter Thirteen - Dataset insight and variable influences established using correlations, regressions, and transparent customized formula optimization, in: Wood, D.A., Cai, J. (Eds.), *Sustainable Geoscience for Natural Gas Subsurface Systems, The Fundamentals and Sustainable Advances in Natural Gas Science and Eng.* Gulf Professional Publishing, pp. 383–408. <https://doi.org/10.1016/B978-0-323-85465-8.00002-9>
- World Health Organisation, 2019. Health consequences of air pollution on populations [WWW Document]. URL <https://www.who.int/news-room/detail/15-11-2019-what-are-health-consequences-of-air-pollution-on-populations> (accessed 4.8.21).
- Wu, W., Dong, B., Wang, Q.R., Kong, M., Yan, D., An, J., Liu, Y., 2020. A novel mobility-based approach to derive urban-scale building occupant profiles and analyze impacts on building energy consumption. *Applied Energy* 278.
<https://doi.org/10.1016/j.apenergy.2020.115656>
- Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. *Knowl Inf Syst* 14, 1–37.
<https://doi.org/10.1007/s10115-007-0114-2>
- Xiao, J., Ye, H., He, X., Zhang, H., Wu, F., Chua, T.-S., 2017. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks.
<https://doi.org/10.48550/arXiv.1708.04617>
- Xu, X., Zou, P.X.W., 2020. Analysis of factors and their hierarchical relationships influencing building energy performance using interpretive structural modelling (ISM) approach. *Journal of Cleaner Production* 272.
<https://doi.org/10.1016/j.jclepro.2020.122650>
- Xu, Z., Shen, D., Nie, T., Kou, Y., 2020. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics* 107, 103465. <https://doi.org/10.1016/j.jbi.2020.103465>

- Yang, L., Liu, S., Liu, J., 2021. The interaction effect of occupant behavior-related factors in office buildings based on the DNAs theory. *Sustainability* (Switzerland) 13. <https://doi.org/10.3390/su13063227>
- Yapa, H., 2001. Window design strategies to maximize the thermal comfort in different climatic zones.
- Yezioro, A., Dong, B., Leite, F., 2008. An applied artificial intelligence approach towards assessing building performance simulation tools. *Energy and Buildings* 40, 612–620. <https://doi.org/10.1016/j.enbuild.2007.04.014>
- Yildiz, B., Bilbao, J.I., Sproul, A.B., 2017. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews* 73, 1104–1122.
- Yin, R.K., 2003. *Case Study Research: Design and Methods*. SAGE.
- Yin, X., Liu, H., Chen, Y., Al-Hussein, M., 2019. Building information modelling for off-site construction: Review and future directions. *Automation in Construction* 101, 72–91. <https://doi.org/10.1016/j.autcon.2019.01.010>
- Yin, Y., Jang-Jaccard, J., Xu, W., Singh, A., Zhu, J., Sabrina, F., Kwak, J., 2023. IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. *Journal of Big Data* 10, 15. <https://doi.org/10.1186/s40537-023-00694-8>
- Yoshino, H., Hong, T., Nord, N., 2017. IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods. *Energy and Buildings* 152, 124–136. <https://doi.org/10.1016/j.enbuild.2017.07.038>
- Yu, F., Qin, Z., Liu, C., Wang, D., Chen, X., 2021. REIN the RobuTS: Robust DNN-Based Image Recognition in Autonomous Driving Systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 40, 1258–1271. <https://doi.org/10.1109/TCAD.2020.3033498>
- Yu, J., 2020. Multiple Angles of Arrival Estimation using Neural Networks. arXiv:2002.00541 [cs, eess].
- Yu, S., Cui, Y., Xu, X., Feng, G., 2015. Impact of Civil Envelope on Energy Consumption based on EnergyPlus. *Procedia Engineering*, The 9th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC) joint with the 3rd International Conference on Building Energy and Environment (COBEE), 12-15 July 2015, Tianjin, China 121, 1528–1534. <https://doi.org/10.1016/j.proeng.2015.09.130>

- Yu, S., Hao, S., Mu, J., Tian, D., 2022. Optimization of Wall Thickness Based on a Comprehensive Evaluation Index of Thermal Mass and Insulation. *Sustainability* (Switzerland) 14. <https://doi.org/10.3390/su14031143>
- Yu, Z., Haghigat, F., Fung, B.C.M., Yoshino, H., 2010. A decision tree method for building energy demand modeling. *Energy and Buildings* 42, 1637–1646.
<https://doi.org/10.1016/j.enbuild.2010.04.006>
- Yuan, J., Nian, V., Su, B., Meng, Q., 2017. A simultaneous calibration and parameter ranking method for building energy models. *Applied Energy* 206, 657–666.
<https://doi.org/10.1016/j.apenergy.2017.08.220>
- Yuan, Y., Shim, J., Lee, S., Song, D., Kim, J., 2020. Prediction for overheating risk based on deep learning in a zero energy building. *Sustainability (Switzerland)* 12, 1–20.
<https://doi.org/10.3390/su12218974>
- Yun, G.Y., Steemers, K., 2011. Behavioural, physical and socio-economic factors in household cooling energy consumption. *Applied Energy* 88, 2191–2200.
<https://doi.org/10.1016/j.apenergy.2011.01.010>
- Zaidan, E., Ghofrani, A., Dokaj, E., 2021. Analysis of Human-Building Interactions in Office Environments: to What Extent Energy Saving Boundaries can be Displaced? *Frontiers in Energy Research* 9. <https://doi.org/10.3389/fenrg.2021.715478>
- Zeng, M., Zou, B., Wei, F., Liu, X., Wang, L., 2016. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data, in: 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS). Presented at the 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), pp. 225–228.
<https://doi.org/10.1109/ICOACS.2016.7563084>
- Zeng, X.-D., Chao, S., Wong, F., 2010. Optimization of bagging classifiers based on SBCB algorithm, in: 2010 International Conference on Machine Learning and Cybernetics. Presented at the 2010 International Conference on Machine Learning and Cybernetics, pp. 262–267. <https://doi.org/10.1109/ICMLC.2010.5581054>
- Zhang, C., Cao, L., Romagnoli, A., 2018. On the feature engineering of building energy data mining. *Sustainable Cities and Society* 39, 508–518.
<https://doi.org/10.1016/j.scs.2018.02.016>
- Zhang, C., Li, J., Zhao, Y., Li, T., Chen, Q., Zhang, X., Qiu, W., 2021. Problem of data imbalance in building energy load prediction: Concept, influence, and solution. *Applied Energy* 297, 117139. <https://doi.org/10.1016/j.apenergy.2021.117139>

- Zhang, C.-X., Zhang, J.-S., 2009. A novel method for constructing ensemble classifiers. *Stat Comput* 19, 317–327. <https://doi.org/10.1007/s11222-008-9094-7>
- Zhang, G., Tian, C., Li, C., Zhang, J.J., Zuo, W., 2020. Accurate forecasting of building energy consumption via a novel ensembled deep learning method considering the cyclic feature. *Energy* 201. <https://doi.org/10.1016/j.energy.2020.117531>
- Zhang, H., Pan, Y., Wang, L., 2017. Influence of plan shapes on annual energy consumption of residential buildings. *International Journal of Sustainable Development and Planning* 12, 1178–1191. <https://doi.org/10.2495/SDP-V12-N7-1178-1191>
- Zhang, J.-P., Li, Z.-W., Yang, J., 2005. A parallel SVM training algorithm on large-scale classification problems, in: 2005 International Conference on Machine Learning and Cybernetics. Presented at the 2005 International Conference on Machine Learning and Cybernetics, pp. 1637-1641 Vol. 3.
<https://doi.org/10.1109/ICMLC.2005.1527207>
- Zhang, Liqiang, Jiang, Y., Li, Y., Zhou, A.J., Cao, J., Liu, S., Wang, Y., Xiao, Z., 2021. Assessment of county-level poverty alleviation progress by deep learning and satellite observations. <https://doi.org/10.21203/rs.3.rs-155105/v1>
- Zhang, L., Wen, J., 2019a. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy and Buildings* 183, 428–442. <https://doi.org/10.1016/j.enbuild.2018.11.010>
- Zhang, L., Wen, J., 2019b. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy and Buildings* 183, 428–442. <https://doi.org/10.1016/j.enbuild.2018.11.010>
- Zhang, Liang, Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., Livingood, W., 2021. A review of machine learning in building load prediction. *Applied Energy* 285, 116452.
<https://doi.org/10.1016/j.apenergy.2021.116452>
- Zhang, L.Y., Jin, L.W., Wang, Z.N., Zhang, J.Y., Liu, X., Zhang, L.H., 2017. Effects of wall configuration on building energy performance subject to different climatic zones of China. *Applied Energy* 185, 1565–1573.
<https://doi.org/10.1016/j.apenergy.2015.10.086>
- Zhang, Q., Lin, Z., Zhang, Haiyan, Bao, X., Zhang, Huxiang, 2020. Prediction of overall survival time in patients with colon adenocarcinoma using DNA methylation profiling of long non-coding RNAs. *Oncology Letters* 19, 1496–1504.
<https://doi.org/10.3892/ol.2019.11236>

- Zhang, X., 2021. Application of data mining and machine learning in management accounting information system. *J. Appl. Sci. Eng.* 24, 813–820.
[https://doi.org/10.6180/jase.202110_24\(5\).0018](https://doi.org/10.6180/jase.202110_24(5).0018)
- Zhang, X., Wang, J., Gao, Y., 2019. A hybrid short-term electricity price forecasting framework: Cuckoo search-based feature selection with singular spectrum analysis and SVM. *Energy Economics* 81, 899–913.
<https://doi.org/10.1016/j.eneco.2019.05.026>
- Zhang, X., Xie, Q., Song, M., 2021. Measuring the impact of novelty, bibliometric, and academic-network factors on citation count using a neural network. *Journal of Informetrics* 15, 101140. <https://doi.org/10.1016/j.joi.2021.101140>
- Zhao, Hai-xiang, Magoulès, F., 2012a. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews* 16, 3586–3592.
<https://doi.org/10.1016/j.rser.2012.02.049>
- Zhao, Hai-xiang, Magoulès, F., 2012b. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews* 16, 3586–3592.
<https://doi.org/10.1016/j.rser.2012.02.049>
- Zhao, Hai-xiang, Magoulès, F., 2012c. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews* 16, 3586–3592.
<https://doi.org/10.1016/j.rser.2012.02.049>
- Zhao, HaiXiang, Magoulès, F., 2012. Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method. *Journal of Algorithms & Computational Technology* 6, 59–77. <https://doi.org/10.1260/1748-3018.6.1.59>
- Zheng, B., Qiu, Y., Aghaei, F., Mirniaharikandehei, S., Heidari, M., Danala, G., 2019. Developing global image feature analysis models to predict cancer risk and prognosis. *Visual Computing for Industry, Biomedicine, and Art* 2, 17.
<https://doi.org/10.1186/s42492-019-0026-5>
- Zhong, H., Wang, J., Jia, H., Mu, Y., Lv, S., 2019. Vector field-based support vector regression for building energy consumption prediction. *Applied Energy* 242, 403–414. <https://doi.org/10.1016/j.apenergy.2019.03.078>
- Zhu, X., Gao, B., Yang, X., Yuan, Y., Ni, J., 2021. Interactions between the built environment and the energy-related behaviors of occupants in government office buildings. *Sustainability (Switzerland)* 13. <https://doi.org/10.3390/su131910607>
- Zhu, Y., 2006. Applying computer-based simulation to energy auditing: A case study. *Energy and Buildings* 38, 421–428. <https://doi.org/10.1016/j.enbuild.2005.07.007>

APPENDIX A: LIST OF RESEARCH PUBLICATIONS

Some chapters of this thesis have been fully or partially published in the following academic Journals between June 2021 to May 2024:

[IF = Impact Factor; SC = CiteScore SJR: Scimago journal ranking]

Journal Papers (*published or accepted for publication*)

1. **Olu-Ajayi, R.**, Alaka, H., Sulaimon, I., Sunmola, F., Ajayi, S., 2022a. Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering* 45, 103406. <https://doi.org/10.1016/j.jobe.2021.103406> [IF: 6.4; SC: 8.3; SJR: Q1]
2. **Olu-Ajayi, R.**, Alaka, H., Sulaimon, I., Sunmola, F., Ajayi, S., 2022b. Machine learning for energy performance prediction at the design stage of buildings. *Energy for Sustainable Development* 66, 12–25. <https://doi.org/10.1016/j.esd.2021.11.002> [IF: 5.5; SC: 7.6; SJR: Q1]
3. **Olu-Ajayi, R.**; Alaka, H.; Owolabi, H.; Akanbi, L.; Ganiyu, S. Data-Driven Tools for Building Energy Consumption Prediction: A Review. *Energies* 2023, 16, 2574. <https://doi.org/10.3390/en16062574> [IF: 3.2; SC: 5.5; SJR: Q2]
4. **Olu-Ajayi, R.**, Alaka, H., Sulaimon, I., Balogun, H., Wusu, G., Yusuf, W., Adegoke, M., 2023. Building energy performance prediction: a reliability analysis and evaluation of feature selection methods. *Expert Systems with Applications* 120109. <https://doi.org/10.1016/j.eswa.2023.120109> [IF: 8.5; SC: 12.6; SJR: Q1]
5. **Olu-Ajayi, R.**; Alaka, H.; Egwim, C.; Grishikashvili, K.; Comprehensive Analysis of Influencing Factors on Building Energy Performance and Strategic Insights for Sustainable Development: A systematic literature review. *Sustainability* 2024. <https://doi.org/10.3390/su16125170> [IF: 3.9; SC: 6.8; SJR: Q2]
6. **Olu-Ajayi, R.**, Alaka, H., Sunmola, F., Ajayi, S. and Mporas, I., "Statistical and Artificial Intelligence-Based Tools for Building Energy Prediction: A Systematic Literature Review," in *IEEE Transactions on Engineering Management*, vol. 71, pp. 14733-14753, 2024, doi: 10.1109/TEM.2024.3422821. [IF: 5.8; SC: 7.6; SJR:Q1]

7. Adegoke, M., Hafiz, A., Ajayi, S., **Olu-Ajayi, R.**, 2022. Application of Multilayer Extreme Learning Machine for Efficient Building Energy Prediction. *Energies* 15, 9512. <https://doi.org/10.3390/en15249512> [IF: 3.2; SC: 5.5; SJR: Q2]
8. Sulaimon, I.A., Alaka, H., **Olu-Ajayi, R.**, Ahmad, M., Ajayi, S., Hye, A., 2022. Effect of traffic data set on various machine-learning algorithms when forecasting air quality. *Journal of Engineering, Design and Technology*. <https://doi.org/10.1108/JEDT-10-2021-0554> [IF: 2.5 SC: 2.7; SJR: Q2]
9. Egwim, C.N.; Alaka, H.; Demir, E.; Balogun, H.; **Olu-Ajayi, R.**; Sulaimon, I.; Wusu, G.; Yusuf, W.; Muideen, A.A. Artificial Intelligence in the Construction Industry: A Systematic Review of the Entire Construction Value Chain Lifecycle. *Energies* 2024, 17, 182. <https://doi.org/10.3390/en17010182> [IF: 3.2; SC: 5.5; SJR: Q2]
10. Balogun, H., Alaka, H., Demir, E., Egwim, C.N., Olu-Ajayi, R., Sulaimon, I. and Oseghale, R., 2024. Artificial intelligence for deconstruction: Current state, challenges, and opportunities. *Automation in Construction*, 166, p.105641. Available at: <https://doi.org/10.1016/j.autcon.2024.105641>. [IF: 9.6; SC: 19.2; SJR: Q1]

Conference Papers (accepted for publication)

1. **Olu-Ajayi, R.**, Alaka, H., 2021. Building energy consumption prediction using deep learning. Environmental Design and Management Conference (EDMIC).
2. **Olu-Ajayi, R.**, Alaka, H., Sulaimon, I., Grishikashvili, K., Sunmola, F., Oseghale, R., Ajayi, S., 2021. Ensemble learning for energy performance prediction of residential buildings. Environmental Design and Management Conference (EDMIC).

APPENDIX B: REVERSE ENGINEERED SYSTEM

```
import numpy as np
from scipy.optimize import minimize

def adjust_features_with_fixed_area(MAINHEAT_DESCRIPTION, TOTAL_FLOOR_AREA,
WALLS_ENERGY_EFF, WALLS_ENV_EFF, ROOF_ENERGY_EFF, ROOF_ENV_EFF, MAINHEAT_ENV_EFF,
FLOOR_DESCRIPTION, FLOOR_HEIGHT, AVG_TEMP, y_given, feature_to_fix_name):
    """
    Adjusts the features to produce the desired target value while fixing a
    specified feature.

    Parameters:
        MAINHEAT_DESCRIPTION (float): Value of the MAINHEAT_DESCRIPTION feature.
        TOTAL_FLOOR_AREA (float): Value of the TOTAL_FLOOR_AREA feature.
        WALLS_ENERGY_EFF (float): Value of the WALLS_ENERGY_EFF feature.
        WALLS_ENV_EFF (float): Value of the WALLS_ENV_EFF feature.
        ROOF_ENERGY_EFF (float): Value of the ROOF_ENERGY_EFF feature.
        ROOF_ENV_EFF (float): Value of the ROOF_ENV_EFF feature.
        MAINHEAT_ENV_EFF (float): Value of the MAINHEAT_ENV_EFF feature.
        FLOOR_DESCRIPTION (float): Value of the FLOOR_DESCRIPTION feature.
        FLOOR_HEIGHT (float): Value of the FLOOR_HEIGHT feature.
        AVG_TEMP (float): Value of the AVG_TEMP feature.
        y_given (float): Desired target value.
        feature_to_fix_name (str): Name of the feature to be fixed.
        model (object): Trained regression model.

    Returns:
        ndarray: Adjusted feature values.
    """
    # Create X_given array
    X_given = np.array([[MAINHEAT_DESCRIPTION, TOTAL_FLOOR_AREA, WALLS_ENERGY_EFF,
WALLS_ENV_EFF, ROOF_ENERGY_EFF, ROOF_ENV_EFF, MAINHEAT_ENV_EFF, FLOOR_DESCRIPTION,
FLOOR_HEIGHT, AVG_TEMP]])

    # Get the index of the specified feature name
    feature_names = ['MAINHEAT_DESCRIPTION', 'TOTAL_FLOOR_AREA',
'WALLS_ENERGY_EFF', 'WALLS_ENV_EFF', 'ROOF_ENERGY_EFF',
'ROOF_ENV_EFF', 'MAINHEAT_ENV_EFF', 'FLOOR_DESCRIPTION',
'FLOOR_HEIGHT', 'AVG_TEMP']
    feature_to_fix_index = feature_names.index(feature_to_fix_name)

    y_pred = loaded_model.predict(X_given.reshape(1, -1))
```

```

result = "This is the predicted energy consumption value: {} and the optimal
values to achieve target energy consumption are listed below:".format(y_pred)

# Get the coefficients (weights) and bias (intercept) from the loaded model
weights = loaded_model.coef_
bias = loaded_model.intercept_

# Define a function to minimize the difference between the predicted and
desired target value
def objective(X):
    X_with_fixed_area = np.copy(X_given)
    X_with_fixed_area[:, feature_to_fix_index] = X_given[:, feature_to_fix_index] # Fix the specified feature
    X_with_fixed_area[:, np.arange(len(X_given[0])) != feature_to_fix_index] = X.reshape(1, -1) # Set the other features to the adjusted values
    return np.abs(loaded_model.predict(X_with_fixed_area) - y_given)

# Adjust the features to produce the desired target value
X_optimal = minimize(objective, X_given[:, np.arange(len(X_given[0])) != feature_to_fix_index].flatten(), method='Nelder-Mead').x
    # Create a DataFrame to associate feature names with optimal values
feature_names.remove(feature_to_fix_name)
optimal_features_df = pd.DataFrame({'Feature': feature_names, 'Optimal Value': X_optimal})
return result, optimal_features_df

```

APPENDIX C: SAMPLE DATA

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
Building 1	4	0	60	2	0	4	4	4	2	2	9	2	2	6	3	3	3	3	3	1	3	3	2	2.62	10.2	55.9	11.4	1015. 2	234	
Building 2	2	1	61	2	0	3	3	0	4	3	3	9	2	2	7	3	3	3	3	3	2	4	4	3	2.37	11.6	66.3	13.4	1016. 4	134
Building 3	4	0	80	2	0	4	4	5	4	2	2	8	1	1	5	3	3	3	3	3	2	4	4	0	2.4	10.9	61.3	12.6	1015. 8	208
Building 4	4	3	138	2	1	6	6	4	4	2	2	30	3	3	7	3	3	3	2	3	0	2	2	3	2.39	10.9	66	14.3	1016. 4	178
Building 5	4	1	147	2	1	6	6	11	4	3	3	30	3	3	5	3	3	3	3	3	2	4	4	0	2.71	10.5	68.5	13.6	1016. 1	120