

Faculty of Computing and Informatics

Department of Informatics

A Hybrid Approach for Electricity Demand Forecasting for the City of Windhoek

Thesis

Submitted in partial fulfilment of the requirements for the degree of

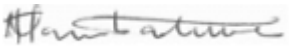
Master of Data Science

at

Namibia University of Science and Technology

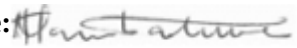
Submitted by:	Ndakalako Diana Twazimina Nambahu
Student Number:	214023419
Supervisor:	Dr Richard Maliwatu
Date of Submission:	26 February 2025

METADATA

Study Title	A Hybrid Approach for Electricity Demand Forecasting for the City of Windhoek
Student Surname, Name	Nambahu Ndakalako Diana Twazimina
Student Number	214023419
Faculty	Faculty of Computing and Informatics
Department	Department of Informatics
Programme	Master of Data Science (09MADS)
Status	Dissertation
Supervisor/Affiliation	Dr Richard Maliwatu
First Co-Supervisor/Affiliation	N/A
Second Co-Supervisor/Affiliation	N/A
Digital Signature of Student	
Digital Signature of Supervisor	
Digital Signature of Programme Coordinator (or equivalent)	
Digital Signature of FCI HDC Representative	

DECLARATION

I, **Ndakalako Diana Twazimina Nambahu**, student number 214023419, hereby declare that the work presented in this dissertation or the degree of Master of Data Science, titled: "**Hybrid Approach for Electricity Demand Forecasting for the City of Windhoek**" is entirely my own. All sources and contributions from other researchers have been duly acknowledged through appropriate references. This work has not been previously submitted to the Namibia University of Science and Technology or any other Institution for academic examination towards any qualification.

Signature:  **Date:** 26 February 2025

ABSTRACT

The accurate forecasting of electricity demand is essential for effective energy management, urban planning, and sustainable resource allocation. Traditional statistical methods and modern machine learning techniques are integrated to leverage their respective strengths, thereby enhancing both accuracy and interpretability of forecasts. This study presents a hybrid approach for electricity demand forecasting.

The research explored electricity consumption trends over the past decade (2014-2023) for the City of Windhoek, through analysing the impact of economic growth, population dynamics, and renewable energy integration by examining a comprehensive dataset that includes historical electricity consumption and factors such as population and economic indicators. Additionally, the study analysed the variance between the City's *notified maximum demand* (NMD) and actual electricity demand (AD) using the City of Windhoek (CoW) historical data and assessed the financial implications of discrepancies. The analysis shows that the City frequently overestimates its NMD, resulting in significant financial losses due to paying for unused capacity. Although the penalty charges for exceeding NMD are relatively minor, the study shows that positive variance, where AD is below the NMD leads to substantial long-term costs.

Six (6) forecasting models, namely *Autoregressive Integrated Moving Average* (ARIMA), *Seasonal Autoregressive Integrated Moving Average* (SARIMA), *Long Short-Term Memory* (LSTM), *Random Forest* (RF), *Gradient Boosting Regression* (GBR), and *LightGBM*, were evaluated for their performance in forecasting electricity demand. Among these, Random Forest and ARIMA models performed superiorly across multiple metrics, namely, Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and R^2 score. These best-performing models were then hybridised using a *stacking ensemble* approach, with Gradient Boosting Regressor (GBR) as the meta-learner, resulting in a hybrid model that outperformed all individual models in terms of forecast accuracy.

Based on the comparative analysis, hybrid models were found to be more accurate than individual models in terms of MAPE, RMSE and R^2 score. The study demonstrates that the hybrid forecast approach can be employed to minimize the discrepancies between the NMD and AD. Besides the forecasting model performance evaluation, the exploratory analysis of historical data uncovered valuable insights for policymakers and energy managers, thereby aiding in the formulation of strategies for efficient energy consumption and sustainable urban development.

Furthermore, the study highlights the potential of integrating hybrid forecasting models into real-world energy management systems to enhance decision-making processes and improve electricity procurement

efficiency. The findings underscore the importance of adaptive forecasting techniques in mitigating financial risks associated with demand estimation errors.

Keywords: Electricity demand forecasting, Machine learning, Statistical, Hybrid approach, Renewable energy

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to the Almighty God for granting me the strength, courage, and wisdom to persevere throughout my studies and during the execution of this thesis.

It is a great honour and privilege to acknowledge the sacrifice, love, dedication, commitment, and unwavering support of these key individuals, all of whom accompanied me on this journey towards this esteemed accomplishment. The order in which they are mentioned does not reflect the depth of my gratitude.

- My dedicated and focused supervisor, Dr Richard Maliwatu, for his continued support and guidance. His motivation and immense knowledge, particularly in the fields of Machine Learning, were crucial elements in carrying out this thesis.

- The course coordinator, Prof. Hippolyte N. Muyingi, for his guidance and assistance with thesis resources and activities.

- The City of Windhoek team, particularly Mr Conrad Witbooi, for assisting with the cumbersome data collection process needed for this study.

- My survey respondents from different energy sector institutions, namely City of Windhoek, NamPower, Cenored, Erongo RED, Nored Electricity, Electricity Control Board (ECB) and others. A big gratitude for sharing your knowledge which was essential in achieving the deliverables of this study.

- My lovely parents and siblings for their unwavering support and for giving me the time and space to focus on my studies when they needed me most.

- My beloved friends and classmates, Angula, Julia, Teopolina, and Dorian, for making this journey memorable and easier, and for turning the hard, sleepless nights into fun experiences.

- All those who contributed directly or indirectly towards the accomplishment of this worthwhile milestone - you are always remembered - friends, relatives, and colleagues.

DEDICATION

This thesis is dedicated to my partner and my parents, whose unwavering love, support, and encouragement have sustained me throughout my academic journey. Their faith in my abilities has been a constant source of strength and motivation.

TABLE OF CONTENTS

METADATA i

DECLARATION ii

ABSTRACT iii

ACKNOWLEDGEMENTS v

DEDICATION vi

TABLE OF CONTENTS vii

LIST OF FIGURES xi

LIST OF TABLES xiii

LIST OF ACRONMYS AND ABBREVIATIONS xiv

CHAPTER ONE: INTRODUCTION 1

1.1 CHAPTER OVERVIEW 1

1.2 BACKGROUND 2

1.3 PROBLEM STATEMENT 3

1.4 RESEARCH AIM, OBJECTIVES AND QUESTIONS 4

1.4.1 Research aim 4

1.4.2 Research objectives 4

1.4.3 Research questions 5

1.4.4 Summary of the findings 6

1.5 SIGNIFICANCE OF THE STUDY 7

1.6 DELIMITATION AND SCOPE OF THE STUDY 8

1.7 ORGANISATION OF THE DISSERTATION 8

CHAPTER TWO: LITERATURE REVIEW 10

2.1 CHAPTER OVERVIEW 10

2.2 LITERATURE REVIEW METHODOLOGY OVERVIEW 11

2.2.1 Search strategy 11

2.2.2 Selection of studies 11
2.2.3 Inclusion/Exclusion criteria 12
2.3 DEFINING ELECTRICITY DEMAND AND FORECASTING 13
2.4 TRADITIONAL STATISTICAL MODELS FOR FORECASTING ELECTRICITY DEMAND 14
2.4.1 ARIMA 14
2.4.2 SARIMA 15
2.4.3 SARIMAX 16
2.5 MACHINE LEARNING MODELS FOR FORECASTING ELECTRICITY DEMAND 18
2.5.1 Random Forest 18
2.5.2 LSTM 18
2.5.3 ANN 19
2.5.4 SVR 19
2.5.5 GBR 19
2.5.6 LightGBM 19
2.6 HYBRID MODELS FOR FORECASTING ELECTRICITY DEMAND 20
2.7 CURRENT FORECASTING METHODS USED IN NAMIBIA AND RELATED COUNTRIES 21
2.8 ANALYSIS OF KEY FACTORS INFLUENCING ELECTRICITY DEMAND 22
2.8.1 Weather conditions 22
2.8.2 Economic and demographic shifts 22
2.8.3 Renewable energy integration 23
2.9 RELATED WORK AND GAPS IN RELATED WORK 23
2.10 CHAPTER SUMMARY 25
CHAPTER THREE: METHODOLOGY 26
3.1 CHAPTER OVERVIEW 26
3.2 RESEARCH PHILOSOPHY 27
3.3 RESEARCH APPROACH 27
3.4 METHODOLOGICAL CHOICE 28
3.5 RESEARCH STRATEGY 28
3.6 RESEARCH TIME HORIZON 29
3.7 TECHNIQUES AND PROCEDURES 29
3.7.1 Data collection 30

3.7.2 Data cleaning and pre-processing: 34
3.7.3 Data visualisation: 41
3.7.4 Modelling 41
3.7.5 Model evaluation: 47
3.7.6 Model hybridisation 49
3.7.7 Electricity consumption forecast 51
3.7.8 Results analysis 53
3.8 CHAPTER SUMMARY 54
CHAPTER FOUR 56
RESULTS AND DISCUSSION OF FINDINGS 56
4.1 CHAPTER OVERVIEW 56
4.2 TRENDS IN ELECTRICITY DEMAND 57
4.2.1 Descriptive statistical analysis 57
4.2.2 Visualisation plots and analysis 59
4.3 CORRELATION BETWEEN DEMAND AND INFLUENCING FACTORS 65
4.3.1 Electricity demand and related variables trends (2014-2023) analysis 65
4.3.2 Correlation of the different variables analysis 67
4.4 **VARIANCE BETWEEN NOTIFIED MAXIMUM DEMAND (NMD) AND THE ACTUAL DEMAND (AD)** 68
4.4.1 Notified Maximum Demand vs Actual Demand vs Ideal Notified Maximum Demand 68
4.4.2 Implications of the variance between NMD and the AD 70
4.4.3 Financial losses due to higher notified maximum demand 71
4.4.4 Demand penalty charges analysis 73
4.4.5 Overall financial implications of maximum demand discrepancies 74
4.5 MODEL PERFORMANCE EVALUATION 75
4.5.1 Hyperparameter tuning analysis 75
4.5.2 ML models evaluation analysis 78
4.5.3 Comparative analysis of selected ML models 82
4.5.4 Statistical models evaluation analysis 85
4.5.5 Statistical models comparisons 87
4.5.6 Hybrid model evaluation analysis 88
4.5.7 Hybrid models comparison 90
4.5.8 All models evaluations 91

4.6 ELECTRICITY CONSUMPTION FORECAST 92
4.7 MODEL VALIDATION 94
4.8 SURVEY OUTCOME RESULTS ANALYSIS 95
4.8.1 Demographic overview 96
4.8.2 Trends in electricity demand 96
4.8.3 Correlation between demand and influencing factors 96
4.8.4 Implications of the variance between NMD and AD 97
4.8.5 Survey and historical data 97
4.9 CHAPTER SUMMARY 98
CHAPTER FIVE: CONCLUSIONS 100
5.1 CHAPTER OVERVIEW 100
5.2 REVISITING THE RESEARCH QUESTIONS 101
5.3 STUDY'S LIMITATIONS 104
5.4 CONTRIBUTION OF THE STUDY TO THE BODY OF KNOWLEDGE 104
5.5 DIRECTIONS FOR FUTURE RESEARCH WORK 105
5.6 CONCLUDING REMARKS 105
REFERENCES 107
APPENDIX A: ETHICAL CLEARANCE CERTIFICATE 113
APPENDIX B: SURVEY QUESTIONS AND RESULTS 114
APPENDIX C: NOTES ON RAW DATA PRE-PROCESSING 119
APPENDIX D: IMPLEMENTATION CODE SNIPPETS 122
APPENDIX E: A REFLECTION ON MY RESEARCH JOURNEY 132

LIST OF FIGURES

Figure 1. 1: Chapter 1 outline map (Source: author) 1
Figure 1. 2: Representation of the electricity supply chain. Source: Teacha (2023) 3

Figure 2. 1: Chapter 2 outline map (Source: author) 10

Figure 3. 1: Chapter 3 outline map (Source: author) 26
Figure 3. 2: Model of the Research process, Source: (Saunders, Lewis, & Thornhill, 2019)..... 27
Figure 3. 3: Data collected and sources (Source: author) 33
Figure 3. 4: Data pre-processing flowchart (Source: author)..... 40
Figure 3. 5: Stationarity check code snippet (Source: author) 43
Figure 3. 6: Differenced code snippet (Source: author)..... 44
Figure 3. 7: ACF and PACF plots (Source: author)..... 45
Figure 3. 8: Training and validation loss (Source: author) 46
Figure 3. 9: Model development process (Source: author)..... 52

Figure 4. 1: Chapter 4 outline map (Source: author) 56
Figure 4. 2: Descriptive statistical analysis (Source: author)..... 57
Figure 4. 3: Distribution of total energy consumption (Source: author)..... 59
Figure 4. 4: Total energy consumption trends (Source: author) 60
Figure 4. 5: Time of use electricity consumption trends (Source: author)..... 62
Figure 4. 6: Box plot of total energy consumed each month (Source: author) 63
Figure 4. 7: Seasonal decomposition of total energy consumption time series (Source: author) 64
Figure 4. 8: Trends of electricity demand and related variables (Source: author)..... 66
Figure 4. 9: Correlation matrix of the different variables (Source: author)..... 67
Figure 4. 10: PCA feature correlations bar plot (Source: author)..... 67
Figure 4. 11: NMD vs AD vs Ideal NMD (Source: author)..... 69
Figure 4. 12: Variance between NMD and AD (Source: author) 70
Figure 4. 13: Financial losses when the NMD is higher than the AD (Source: author) 72
Figure 4. 14: Demand penalty charges over time (Source: author) 73
Figure 4. 15: ML models performance before hyperparameter tuning (Source: author)..... 75
Figure 4. 16: ML models performance after hyperparameter tuning (Source: author)..... 76
Figure 4. 17: Statistical models performance after hyperparameter tuning (Source: author) 77

Figure 4. 18: Actual vs predicted electricity demand for all ML models (Source: author) 82
Figure 4. 19: All ML models evaluation results bar plots (Source: author)..... 83
Figure 4. 20: Actual vs predicted line plot for statistical models (Source: author)..... 87
Figure 4. 21: Statistical models evaluation results bar plots (Source: author)..... 88
Figure 4. 22: Hybrid models evaluation results bar plots (Source: author) 89
Figure 4. 23: Hybrid models actual vs predicted line plot (Source: author) 89
Figure 4. 24: Hybrid model consumption forecast for the next two years (Source: author)..... 92
Figure 4. 25: Electricity consumption forecast numerical values data frame (Source: author) 93
Figure 4. 26: Model validation actual vs predicted electricity consumption (Source: author) 94
Figure 4. 27: Actual energy consumption (kWh) vs hybrid model forecast (Source: author) 95
Figure 4. 28: Hybrid model performance evaluation (Source: author)..... 95

Figure 5. 1: Chapter 5 outline map (Source: author) 100

LIST OF TABLES

Table 1. 1: Organisation of the dissertation 8

Table 2. 1: Literature review inclusion criteria 12

Table 3. 1: CoW historical electricity consumption dataset selected features 31
Table 3. 2: Economic indicators and population dataset features 32
Table 3. 3: Pre-processing of CoW historical energy consumption raw data 35
Table 3. 4: Pre-processing Economic, Population and Renewable Energy Raw Data 37
Table 3. 5: ML models training methodology 45
Table 3. 6: Statistical models training methodology 46
Table 3. 7: Evaluation metrics 47
Table 3. 8: Description of the meta-learner models used to hybridise 49
Table 3. 9: Summary of research objectives and evaluation criteria 54

Table 4. 1: General descriptive statistical characteristics 58
Table 4. 2: Statistical descriptive analysis 58
Table 4. 3: ML models evaluation analysis 78
Table 4. 4: All ML Models Evaluation Results 82
Table 4. 5: Statistical models evaluation analysis 85
Table 4. 6: Statistical models evaluation results 87
Table 4. 7: Hybrid model evaluation results 88
Table 4. 8: All models evaluation results 91
Table 4. 9: Results and discussion of findings summary 98

LIST OF ACRONMYS AND ABBREVIATIONS

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
ANN	Artificial Neural Network
AR	Autoregression
ARIMA	Autoregressive Integrated Moving Average
BiLSTM	Bidirectional Long Short-Term Memory
CoW	City of Windhoek
CNN	Convolutional Neural Network
DG	Distributed Generation
DNN	Deep Neural Network
DT	Decision Tree
ECB	Electricity Control Board
GBR	Gradient Boosting Regressor
GBRT	Gradient Boosted Regression Trees
GDP	Gross Domestic Product
GRU	Gated Recurrent Unit
HWES	Holt-Winters Exponential Smoothing
kVA	Kilovolt-Amperes
kWh	Kilowatt-Hour
LSTM	Long Short-Term Memory
LR	Linear Regression

LightGBM	Light Gradient Boosting Machine
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
N$	Namibian Dollar
N$/kWh	Namibian Dollar per Kilowatt-Hour
NEF	Namibian Energy Fund
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
PC	Principal Component
R^2	R-Squared
RFE	Recursive Feature Elimination
RF	Random Forest
RFR	Random Forest Regressor
RMSE	Root Mean Squared Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous Factors
STL	Seasonal and Trend Decomposition using Loess
SVR	Support Vector Regressor
VAR	Vector Autoregression
XGBoost	Extreme Gradient Boosting

CHAPTER ONE: INTRODUCTION

1.1 CHAPTER OVERVIEW

This chapter gives an overview to the research by highlighting the primary motivation for the focus on electricity demand forecasting for the City of Windhoek. The chapter presents the problem statement, the conceptual framework and the key research objectives and questions guiding the work, emphasizing the importance and relevance of developing a hybrid approach to enhance electricity demand forecasting accuracy. The map shown in Figure 1. 1 below outlines the chapter layout.

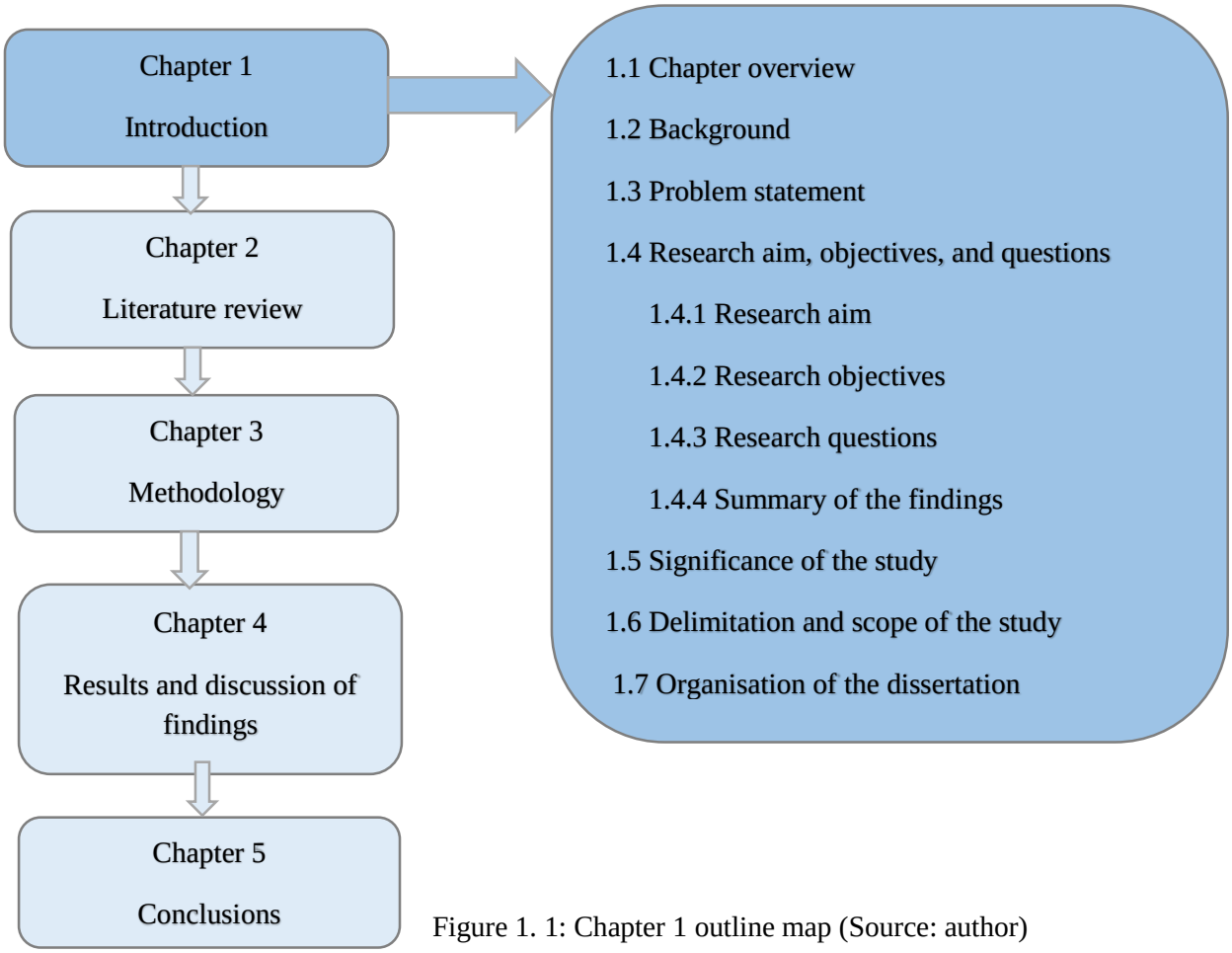


Figure 1. 1: Chapter 1 outline map (Source: author)

1.2 BACKGROUND

Electricity demand represents the average rate at which electricity is consumed by homes, businesses, or utilities within a specific time period. Typically, the demand increases when more electrical devices or machines are used simultaneously (Dixie Electric Power Association, 2024). Forecasting electricity demand is crucial for modern energy management, and it becomes particularly challenging when applied to rapidly growing urban areas like the City of Windhoek in Namibia. Namibia, like many developing economies, experiences fluctuations in electricity demand due to factors such as urban migration, industrial expansion, and renewable energy integration.

Accurate forecasting enables energy providers to efficiently allocate resources, prevent shortages or financial losses, and maintain the stability of the electrical grid. By matching supply with demand, energy providers can minimise the need for costly emergency power generation and reduce waste (Alam & Arefifar, 2019). Moreover, accurate forecasts support infrastructure planning, optimise the incorporation of renewable energy sources, and facilitate demand response strategies that balance energy consumption with availability. Anticipating future demand is vital for ensuring a reliable and continuous power supply, which is essential for economic growth, public safety, and overall quality of life.

Electricity providers typically purchase bulk electricity from power generators and then distribute it to end consumers at the retail level for use in homes and businesses. Different countries adopt varying approaches to this process. Some use a vertically integrated utility model/bundled approach where a single company manages the entire process of electricity generation, transmission, and distribution. Others follow an unbundled approach, where these components are separated into distinct entities responsible for generation, transmission, distribution, and retail.

In the case of Windhoek, the electricity supply chain follows an unbundled structure. End consumers, such as households, businesses, and industries, represent the primary source of energy demand. These consumers interact with the City of Windhoek, the electricity distributor, which functions as a retailer. The City of Windhoek buys electricity in bulk from Nampower, the national power utility responsible for generating and transmitting electricity throughout Namibia. Figure 1. 2 highlights the distinct phases in the electricity supply chain, mirroring how the City of Windhoek interfaces with Nampower. Nampower, as the national power utility, handles the **generation** and **transmission** of electricity across Namibia, while the City of Windhoek focuses on **distribution** and retail, delivering electricity to end consumers within the city.

While Namibia's primary source of electricity generation through NamPower is hydropower, which is renewable, the City of Windhoek’s electricity supply is sourced from the Van Eck coal power plant, which

is non-renewable. Thus, the City of Windhoek additionally focuses on initiatives to reduce electricity costs and integrate renewable energy sources like solar power into its grid (Graig, 2024).

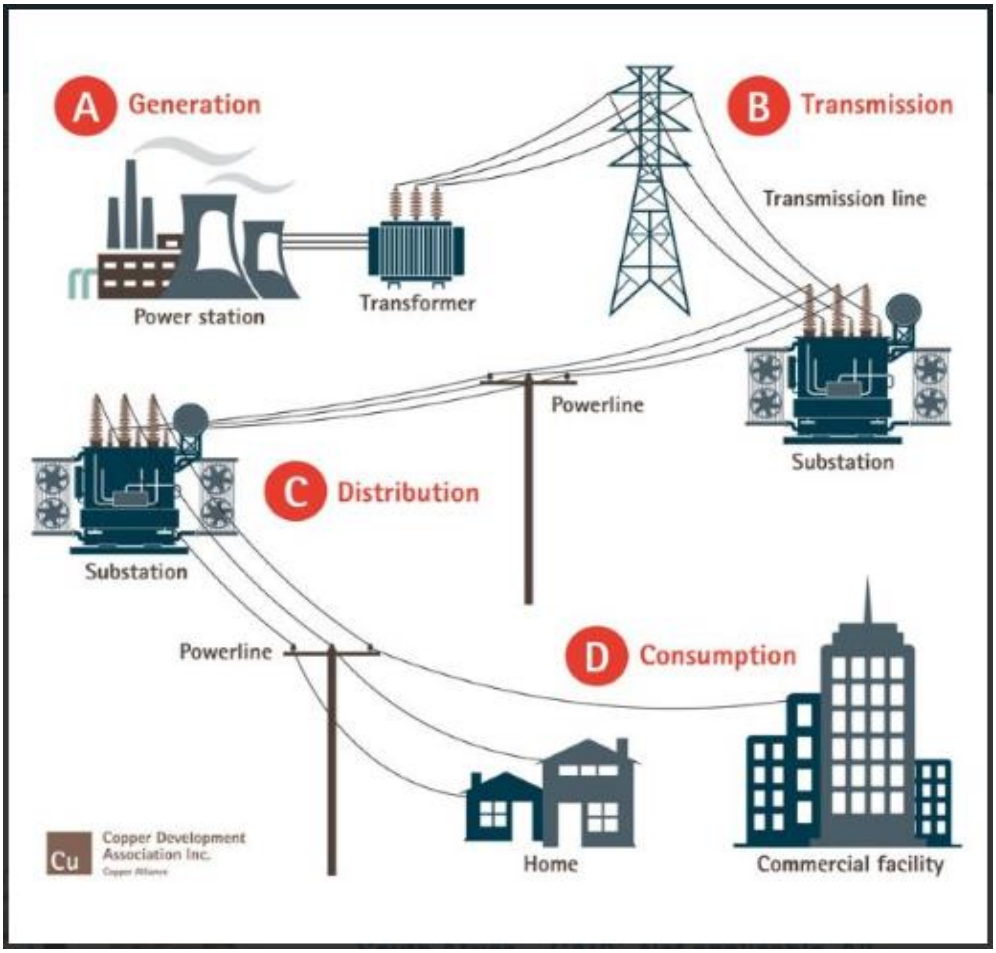


Figure 1. 2: Representation of the electricity supply chain. Source: Teacha (2023)

1.3 PROBLEM STATEMENT

Electricity demand forecasting is vital for effective energy resource allocation, infrastructure development, and sustainable energy management. Understanding load forecasting methods in depth is crucial for businesses and policymakers to better forecast future electricity needs (Mir et al., 2020). The City of Windhoek faces difficulties in accurately forecasting electricity demand due to complex interplay of factors, including the integration of renewable energy into the power grid, economic fluctuations and population growth. Inaccurate forecasts can result in overloads, blackouts, or misallocated

resources, potentially causing substantial economic impacts on the city. Moreover, traditional statistical methods may fail to capture the complex data patterns, while machine-learning approaches may lack interpretability.

One significant challenge is the determination of the Notified Maximum Demand (NMD), which is crucial for ensuring that the electricity supply meets the city's needs without incurring unnecessary costs. According to Nampower (2020), NMD refers to the peak demand level specified in written communication by the City of Windhoek and recognised by Nampower as the maximum supply requirement to be available upon request. Accurately setting the NMD is vital because any discrepancies between the NMD and Actual Demand (AD) can lead to financial penalties for exceeding the demand or unnecessary costs from underutilising the contracted capacity. Considering the density of factors affecting electricity demand - such as population growth, economic trends, and the integration of renewable energy - there is a need for a forecasting approach that can accurately predict electricity demand and inform the appropriate NMD.

This thesis, therefore, aimed to implement a hybrid forecasting approach, specifically utilising a stacking ensemble method, which combines the strengths of both traditional and machine learning techniques. This method seeks to improve the accuracy and interpretability of long-term electricity demand forecasts for Windhoek, thereby supporting the city's energy sustainability goals.

1.4 RESEARCH AIM, OBJECTIVES AND QUESTIONS

1.4.1 Research aim

The study aimed to establish a forecasting method that leverages both machine learning and statistical methods to use historical energy data for accurately forecasting electricity demand in the City of Windhoek. This approach can enable the City of Windhoek and other distribution companies in Namibia to determine the correct Notified Maximum Demand (NMD) to declare to Nampower, thereby avoiding overcharges from overstated demand or fines from exceeding declared demand.

1.4.2 Research objectives

The study's main objective is to apply a hybrid forecasting approach that combines both the traditional time series methods and machine learning algorithms for electricity demand forecasting for the City of Windhoek.

The sub-objectives of the study are to:

1. explore the trends in electricity demand for the City of Windhoek for the past ten years;
2. analyse the relationship between electricity demand and factors thought to influence electricity demand, such as GDP, population growth and renewable/alternative energy integration;
3. analyse the variance between the Notified Maximum Demand (NMD) and the Actual Demand (AD) for CoW, and assess the financial implications of any disparities;
4. evaluate the performance of commonly used machine learning and statistical models to determine the most accurate model for forecasting electricity demand;
5. forecast the monthly electricity demand for the City of Windhoek for the next two years; and
6. analyse energy experts' perceptions alongside historical data to assess key factors impacting the electricity distribution industry.

1.4.3 Research questions

The primary research question driving this study is: "How can a hybrid forecasting approach, combining traditional time series methods and advanced machine learning techniques, enhance the accuracy and interpretability of electricity demand forecasting for the City of Windhoek?"

The sub-research questions of the study are:

1. What are the noticeable trends and patterns in electricity demand over the past decade (2014-2023) in Windhoek?
2. What correlations exist between alternative energy adoption/integration, population growth, and GDP?
3. How does the variance between the City of Windhoek's officially Notified Maximum Demand and the observed actual electricity demand influence the city's financial outcomes, particularly in terms of monthly energy procurement?
4. Which machine learning and statistical model demonstrates the highest accuracy for forecasting electricity demand in Windhoek, and how does the performance of these models compare to the hybrid?
5. What will be the monthly electricity demand for the City of Windhoek for the next two years (April 2024 – March 2026), and how will the demand vary over the months?
6. How do energy experts' perceptions, when analysed alongside historical data, help to identify the key factors influencing the electricity demand?

1.4.4 Summary of the findings

The analysis of electricity demand trends in Windhoek from 2014 to 2023 revealed distinct cyclical patterns, with seasonal fluctuations leading to higher consumption during the winter months (June-August) due to increased heating demands. Over the decade, electricity consumption showed a general relatively stable trend, without a consistent upward nor downward trajectory.

A correlation analysis between GDP, which is often considered the main economic performance indicator, population growth, and renewable energy integration revealed weak positive correlations between GDP and other variables, thus suggesting that GDP growth does not strongly influence nor is it influenced by other factors. Additionally, a strong negative correlation was identified between overall electricity consumption and the incorporation of renewable energy sources, thus suggesting that as renewable energy capacity and population increase, overall electricity consumption decreases, possibly due to improved energy efficiency and a shift toward sustainable practices.

Significant differences between the City of Windhoek's official NMD and the AD were observed. These discrepancies had notable financial implications on the City's monthly energy procurement costs. Before 2018, the AD frequently exceeded the NMD, which resulted in penalty charges of N$ 9 072. 98 on a monthly average. Post-2018, AD continued to fluctuate but way below the NMD, which had been increased to 190000 kVA. When the NMD is higher than the AD, the CoW is obligated to pay for unused electricity capacity, which results in unnecessary expenses. On average, this overestimation cost the City about N$ 1,447,058.00 monthly.

The evaluation of commonly used ML and statistical models for electricity demand forecasting showed that Random Forest and ARIMA outperformed the other models in terms of MAPE, RMSE and R^2 score. However, the hybrid model, which combined these two approaches using a Gradient Boosting Regressor as a meta-learner, outperformed all individual models, thus demonstrating superior accuracy and reliability. This highlights the effectiveness of hybridization in reducing forecast errors and enhancing predictive performance.

The forecast for monthly electricity consumption in Windhoek from April 2024 to March 2026 suggests that the City will experience minor fluctuations but will remain generally stable over the next two years. The consumption peaks are expected around mid-year and at the end of the year, due to seasonal and holiday-related factors.

The combined analysis of energy experts' perceptions and historical data revealed that economic growth, industrial activities, and renewable energy integration are key drivers of electricity demand in Windhoek. Experts highlighted initial set-up costs and regulatory barriers as the major inhibiting factors influencing the adoption of alternative energy sources. The findings emphasise the need for strategic investments in local power generation and grid improvements to enhance energy security and optimise electricity distribution amid increases in electricity demand and uncoordinated adoption of alternative energy sources.

1.5 SIGNIFICANCE OF THE STUDY

The significance of this study is in its ability to improve electricity demand forecasting by effectively combining traditional time series methods with modern machine learning techniques. This approach has resulted in more accurate and interpretable forecasting models, which directly contribute to improved energy management and informed decision-making for the City of Windhoek. By accurately forecasting electricity demand, the study mitigates the financial risks associated with misalignment between the Notified Maximum Demand (NMD) and the actual demand (AD), thereby minimising the potential for overcharges and financial losses due to underutilisation.

From an academic perspective, this study bridges the gap between traditional and modern methodologies, thus advancing the field of energy demand forecasting. Practically, the findings empower the City of Windhoek to optimise energy resource allocation, infrastructure planning, and overall energy management. The analysis also emphasises the financial burden placed on consumers due to rising electricity costs. Recent reports indicate that N$100 now buys fewer than 40 electricity units, thus underscoring the increasing cost of electricity for end-users in Windhoek (The Namibian, 2024). This highlights the importance of accurate demand forecasting to ensure that energy procurement aligns closely with actual consumption, thereby minimizing unnecessary financial strains on both the city and its residents.

By extension, this study contributes to the sustainable management of energy resources, optimizing infrastructure planning, and promoting energy efficiency in urban settings. Furthermore, by incorporating renewable energy trends into the forecasting process, the study supports the transition to a greener energy mix, thereby contributing to a more resilient and environmentally sustainable energy ecosystem.

1.6 DELIMITATION AND SCOPE OF THE STUDY

This study focuses on forecasting electricity demand for the City of Windhoek using historical data on key variables such as energy consumption, GDP, and population. The study is limited to data collected over a ten-year period, from 2014 to 2023. While several factors can influence electricity demand, this study excludes variables such as weather conditions, energy prices, and technological advancements in energy efficiency, as they are outside the scope of the available data.

Methodologically, the study is bounded by the use of a hybrid approach combining traditional statistical models and machine learning techniques. The focus is on models such as ARIMA, SARIMA, LSTM, Random Forest, Gradient Boosting, and LightGBM. The study does not explore other forecasting methods or alternative machine learning models outside this selection.

1.7 ORGANISATION OF THE DISSERTATION

The rest of the dissertation is divided into four additional chapters, namely, Literature Review, Methodology, Results and Discussion of the Findings, and Conclusion. Each chapter serves a distinct purpose and was structured to facilitate a logical flow of information as summarised in Table 1. 1.

Table 1. 1: Organisation of the dissertation

Chapter	Description
2. Literature Review	The literature review chapter presents a summary of current research related to electricity demand forecasting. The chapter synthesises findings from recent studies, highlighting various methodologies, models, and their effectiveness in forecasting electricity consumption.
3. Methodology	The methodology chapter details the approach used in addressing the study's objectives. The chapter outlines the processes of data collection, preparation, model training, and performance evaluation.
5. Results and Discussion of the Findings	The chapter presents the data analysis and discussion of the findings. It includes the performance results for selected models considered in this study. The results are displayed in tables and graphs to facilitate comparison. The findings are interpreted in relation to the research objectives posed in the introduction.

6. Conclusion	In concluding the dissertation, this chapter summarises the study's key findings and discusses their broader implications. The chapter reflects on the hybrid forecast modelling approach and how this went towards addressing the research objectives. The chapter highlights the limitations of the study and discusses future work and further explorations.

CHAPTER TWO: LITERATURE REVIEW

2.1 CHAPTER OVERVIEW

This chapter presents a narrative literature review, drawing on widely consulted scientific literature relevant to the primary focus of this research, as overviewed in Figure 2.1.

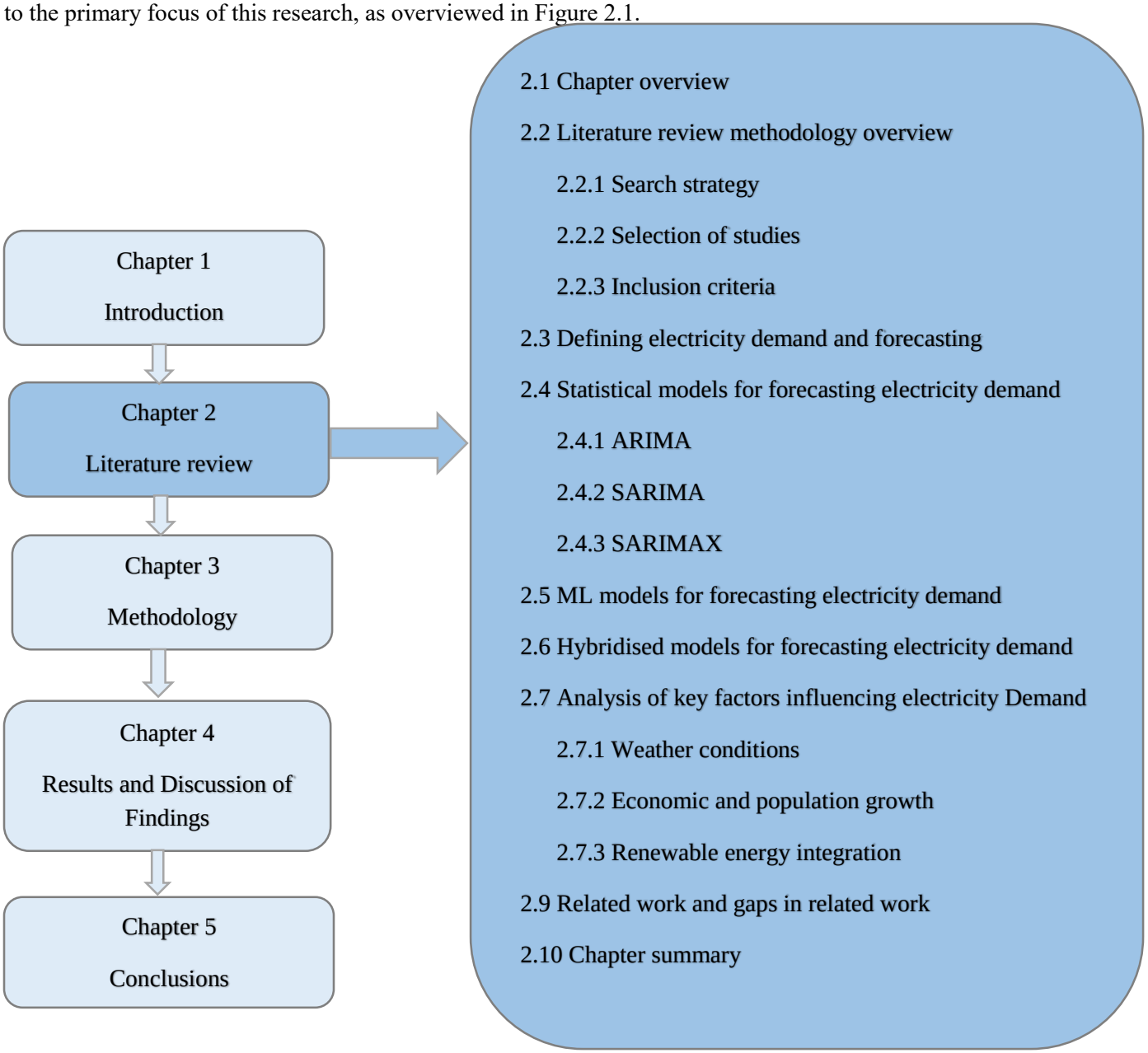


Figure 2. 1: Chapter 2 outline map. Source: author

2.2 LITERATURE REVIEW METHODOLOGY OVERVIEW

This section outlines the approach employed in conducting the literature review, thereby ensuring a comprehensive and methodologically sound selection of studies. The process involved a well-defined search strategy, study selection, and inclusion/exclusion criteria to gather relevant literature that aligns with the research objectives of this study.

2.2.1 Search strategy

A search was conducted across multiple academic databases such as ScienceDirext, IEEE Xplore Digital Library database, ProQuest, SpringerLink, and Web of Science, to identify relevant studies on electricity demand forecasting. The search strings incorporated keywords related to electricity demand forecasting, hybrid models, machine learning, traditional statistical models, and influencing factors such as economic trends, population growth, and renewable energy integration. Relevant publications were manually reviewed to extract essential evidence for analysis. Titles and abstracts of studies deemed potentially relevant were thoroughly examined, and articles were chosen based on their alignment with the predefined search criteria.

The search was conducted in three phases:

1. A **general search** focusing on comprehensive keywords related to electricity demand forecasting;
2. A **hybrid models search** to identify papers specifically addressing the integration of traditional statistical and machine learning models; and
3. A **factor-specific search** focusing on studies that explore key factors influencing electricity demand such as population growth, weather conditions, economic trends, and renewable energy integration.

Logical operators (AND, OR) were used to refine the search, ensuring the retrieval of relevant publications. The search covered literature published from 2019 to 2024, capturing the most recent advancements and seminal works in the field.

2.2.2 Selection of studies

The selection process for studies involved a three-layer filtering approach i.e title, abstract and paper content screening to ensure the inclusion of only the most relevant research. A paper was selected if the title contained relevant keywords such as “electricity demand forecasting,” “machine learning,” “traditional statistical models,” and “hybrid approach”. The abstract was then reviewed to assess whether the study's

objectives and methodologies aligned with the research focus, including the application of traditional statistical or machine learning methods and exploration of key influencing factors. Finally, the full text of the remaining studies was reviewed to ensure they provided detailed insights into forecasting models, offered empirical results, and aligned with the research objectives.

2.2.3 Inclusion/Exclusion criteria

The inclusion criteria for this literature review are summarised in Table 2. 1 below. Only studies that met all three layers of filtering were considered. Specifically, the studies had to be written in English and published as reviewed journal articles, conference papers, or reports between the years 2019-2024. The emphasis was on studies that applied hybrid forecasting models, examined key factors influencing electricity demand, or utilised machine learning techniques, and traditional statistical methods.

Table 2. 1: Literature review inclusion criteria

Criteria	Definition
Manuscript language	English
Accepted paper types	Conference papers, journal publications, reports
Time Period	2019-2024
Focus	Electricity demand forecasting, hybrid models, key influencing factors, machine learning techniques, and traditional statistical methods.

Studies were excluded according to the following criteria: non-English publications, non-peer-reviewed materials, review papers, editorials, and studies lacking full-text availability. Additionally, studies focusing solely on technical aspects of electrical engineering without relevance to demand forecasting were excluded, and studies not aligning with the criteria outlined in Table 2. 1.

2.3 DEFINING ELECTRICITY DEMAND AND FORECASTING

In the recent years, forecasting electricity demand has gained importance as a research area due to its crucial role in daily life and its significant impact on a country's socioeconomic development (Energy Agency, 2024). Accurate forecasting is essential for effective energy management, ensuring that supply meets demand, thus avoiding costly overproduction or blackouts.

In recent years, forecasting electricity demand has gained importance as a research area due to its crucial role in daily life and its significant impact on a country's socioeconomic development.

Electricity demand refers to the total amount of electrical energy consumed by end-users over a specific period, a definition that underscores the need for precise forecasting to support infrastructure investment, generation capacity, and policy-making (Straight, 2023). The study by Román-Portabales et al. (2021) emphasises the importance of clear definitions and methodologies in establishing a reliable framework for demand forecasting. It discussed various forecasting horizons - long-term, medium-term, and short-term - and their implications for energy management strategies. Accurate forecasting is further supported by integrating various factors, which can significantly influence electricity consumption. Short-term forecasting typically covers a period from a few hours to a week, focusing on operational and immediate resource allocation needs (Banimustafa & Al-Omari, 2022). Medium-term forecasting spans from one week to a year and is crucial for tactical planning, such as maintenance scheduling and fuel purchasing. Long-term forecasting, which extends beyond a year, is primarily used for strategic planning, including infrastructure development and policy formulation. As the forecast horizon extends, the accuracy of predictions generally decreases due to the compounding uncertainty of influencing factors.

Upadhaya et al. (2019) further emphasise that electricity demand forecasting is a critical process for predicting future electricity usage based on various factors, including historical data, weather conditions, and economic indicators. The incorporation of these diverse factors into forecasting models is crucial for enhancing the accuracy of predictions, particularly in regions where external influences such as weather can have a significant impact on consumption patterns.

2.4 TRADITIONAL STATISTICAL MODELS FOR FORECASTING ELECTRICITY DEMAND

Before the advent of machine learning methods, traditional electricity forecasting methods primarily relied on straightforward techniques based on historical consumption data. These methods formed the foundation of early electricity demand forecasting, with regression and autoregressive models being particularly prominent. A detailed review by Klyuev et al. (2022) provided a thorough overview of these traditional approaches, highlighting their role in laying the groundwork for more sophisticated forecasting methods. Recent studies have also reinforced the effectiveness of traditional statistical models in forecasting electricity demand across various contexts.

2.4.1 ARIMA

ARIMA is a commonly used statistical model in time series forecasting, particularly in electricity demand forecasting. It captures non-stationary time series data by combining the autoregressive (AR) and moving average (MA) terms with differencing to make the data stationary (Brownlee, 2023). The model is mathematically expressed as a combination of these three components: AR, differencing (I for integration), and (MA), denoted as ARIMA(p, d, q), where p represents the number of lag observations included, d indicates the degree of differencing needed to achieve stationarity, and q denotes the size of the moving average window. This flexibility allows ARIMA to accommodate various patterns within the data, making it highly adaptable to different forecasting environments. The model is mathematical expressed in

Equation form as follows:

$$y_t = c + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{j=1}^{q} \Theta_j \, \epsilon_{t-j} + \epsilon_t$$

Equation 2.1

where:

y_t is the actual value at time t

c is a constant

ϕ_i represents the coefficients of the autoregressive terms

Θ_j denotes the coefficients of moving average terms

ϵ_t is the white noise error term

ARIMA models are commonly used in electricity demand forecasting because they can effectively capture model both short-term fluctuations and long-term trends in time series data. Studies have demonstrated that ARIMA, by incorporating lagged observations and error terms, could produce accurate forecasts for electricity demand, even when dealing with complex time series data. This adaptability is evident in various applications, such as in forecasting real-time load data, where ARIMA has proven effective in predicting short-term fluctuations and responding to changes in demand patterns in study done in southern Tripoli, Libya. However, ARIMA's weakness lies in its assumption of linearity in the time series data, which can be a limitation when dealing with complex real-world data that may exhibit nonlinear patterns. ARIMA models are best suited for time series that are stationary and have a linear relationship among the data points. However, in many cases, electricity demand can be subjective to nonlinear factors such as sudden changes in weather, economic conditions, or policy shifts, which ARIMA may not capture effectively. Additionally, the need for the data to be stationary means that ARIMA may require extensive preprocessing such as differencing or transformation, to make the data suitable for modeling, which can complicate the analysis and potentially lead to loss of important information (Azka et al., 2020; Goswami & Kandali, 2022).

2.4.2 SARIMA

SARIMA extends ARIMA by incorporating seasonal differencing and seasonal AR and MA components to handle seasonality in time series data (Bajaj, 2023), thus making it mainly suitable for electricity demand forecasting where seasonal patterns are prevalent. The SARIMA model is denoted as ARIMA (p, d, q) (P, D, Q, s), where: p, d, and q are the non-seasonal AR, differencing, and MA terms, respectively, and P, D, and Q represent the seasonal AR, differencing, and MA components, respectively with s indicating the length of the seasonal cycle. The model is mathematical expressed in Equation 1 as follows:

$$y_t = c + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{j=1}^{q} \Theta_j \, \epsilon_{t-j} + \sum_{k=1}^{P} \Phi_k \, Y_{t-k} + \sum_{l=1}^{Q} \Theta_l \, \epsilon_{t-l} \, \epsilon_t \qquad \text{Equation 1.2}$$

y_t is the actual value at time t

c is a constant

ϕ_i represents the coefficients of the autoregressive terms

Θ_j denotes the coefficients of moving average terms

Φ_k and Θ_l are the coefficients of the seasonal AR and SMA terms, respectively.

ϵ_t is the white noise error term.

SARIMA is particularly useful in scenarios where the data exhibits regular seasonal patterns such as monthly or quarterly fluctuations in electricity demand. A study by Elsaraiti et al. (2021) validated the effectiveness of SARIMA in forecasting electricity consumption, emphasizing its ability to model both the non-seasonal and seasonal components of the time series. This model was able to capture the cyclical nature of electricity demand, which is often subjective to factors such as weather conditions and seasonal activities. Similarly, Shiwakoti et al. (2023) applied SARIMA to forecast electricity demand, highlighting its superiority over non-seasonal models in terms of capturing seasonal effects and providing more accurate forecasts. However, while SARIMA is effective in handling seasonality, its complexity increases with the addition of seasonal parameters, which can make the model more difficult to implement and interpret in comparison to simpler models like ARIMA. Moreover, SARIMA assumes linearity in the time series data, similar to ARIMA, which may limit its applicability in cases where nonlinear relationships are significant (Goswami & Kandali, 2022).

2.4.3 SARIMAX

SARIMAX is an extension of the Seasonal ARIMA (SARIMA) model that includes exogenous variables (denoted as "X") into the forecasting process, allowing the model to account for external influences on the dependent variable. The model is expressed mathematically in Equation format as:

$$y_t = c + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{j=1}^{q} \Theta_j \, \epsilon_{t-j} + \sum_{k=1}^{r} \beta_k \, X_{t-k} + \epsilon_t + (Seasonal\ terms) \qquad \text{Equation 2.3}$$

where

y_t is the actual value at time t

c is a constant

ϕ_i represents the coefficients of the autoregressive terms

Θ_j denotes the coefficients of moving average terms

β_k represents the coefficients of exogenous variables X_{t-k}

ϵ_t is the white noise error term

The seasonal terms include seasonal autoregressive and moving average components

SARIMAX effectively captures the influence of external factors on the time series by including exogenous variables, which makes it particularly suitable for electricity demand forecasting where external factors such as economic indicators, weather conditions, and other time-varying influences are critical. A study by Cetinkaya and Acarman (2021) utilised SARIMAX for next-day electricity demand forecasting by integrating variables like temperature and economic activity levels. This approach enhanced the model's accuracy, especially in scenarios where electricity demand was significantly impacted by factors beyond its historical patterns and inherent seasonality.

However, a comparative study focusing on forecasting hourly electricity demand, found that Ridge regression, a more advanced regression technique that effectively accounts for hourly and daily seasonality, outperformed SARIMAX. Ridge regression achieved the lowest MAE by optimizing day-ahead market participant portfolios, demonstrating the value of incorporating more sophisticated regression techniques in electricity forecasting.

Overall, these studies collectively highlight the continued relevance and effectiveness of traditional statistical models in electricity demand forecasting. These models excel in environments where the data exhibits linearity, stationarity, and predictable seasonality. However, they tend to struggle in scenarios where the data is highly non-linear, exhibits non-stationarity that cannot be easily corrected through differencing, or is influenced by a multitude of interacting variables that are not easily captured by linear models. In such cases, machine learning models offer distinct advantages because of their ability to automatically identify complex patterns, non-linear relationships, and interactions between variables without requiring extensive manual feature engineering. These models are especially effective for handling large datasets, high-dimensional data, or situations where the underlying data-generating process is not well understood.

The analysis of the surveyed literature on statiscal models applied in electricity demand forecasting revealed that SARIMA and ARIMA emerge as the predominant models. SARIMA, due to its ability to effectively model seasonality in time-series data, has been the preferred choice in scenarios where seasonal patterns play a crucial role in electricity consumption. Similarly, ARIMA, while lacking a seasonal component, continues to be widely used, particularly in cases where electricity demand is less influenced by seasonal factors. Other models such as HWES, Ridge Regression, and SARIMAX, although not as commonly applied as ARIMA AND SARIMA, they are chosen for specific scenarios where their particular strengths are well-suited to the forecasting requirements.

2.5 MACHINE LEARNING MODELS FOR FORECASTING ELECTRICITY DEMAND

Machine learning approaches have emerged as powerful tools that effectively incorporate exogenous factors and calendar related variables such as GDP, population growth and seasonal effects into the forecasting process. By leveraging large datasets and sophisticated algorithms, machine learning models can capture complex patterns and relationships within the data, resulting in more accurate forecasts.

2.5.1 Random Forest

Random Forest is a versatile and robust ensemble learning method that has proven to be particularly effective in electricity demand forecasting. The model operates by constructing a multitude of decision trees during training and produces the average prediction from the individual trees (GeeksforGeeks, 2024). This ensemble approach helps minimize overfitting and enhances the model's ability to generalise, making it suitable for handling large datasets with diverse features.

The effectiveness of Random Forest in electricity demand forecasting is well-documented across various studies. Studies have demonstrated its capability to capture intricate dependencies and patterns in electricity consumption data, which are often influenced by factors such as economic indicators, weather conditions, and population growth. It has been shown to perform accurately in different forecasting horizons, which is critical for effective energy management and planning. The model's ability to dynamically weigh the importance of various features allows it to adapt to changing data patterns, making it particularly effective in regions with diverse demand conditions. Additionally, its effectiveness in regions with fluctuating and unpredictable electricity demand underscores its robustness and versatility as a forecasting tool, thus ensuring accurate and reliable demand predictions even in complex environments (Giamarelos et al., 2023).

2.5.2 LSTM

Long Short-Term Memory (LSTM) networks, a specialised form of recurrent neural networks (RNNs), are particularly effective in handling time series forecasting tasks due to their unique architecture, which includes memory cells and gates (Lindemann et al., 2020). These components allow LSTMs to retain and update information over extended periods, making them exceptionally well-suited for capturing long-term dependencies in sequential data, such as electricity consumption patterns. This capability is crucial for tasks where the future values depend not just on the most recent data points but also on trends and patterns that span across longer periods.

Wong et al. (2022) applied LSTM networks to forecast electricity usage, demonstrating the model's superior ability to capture the temporal dependencies in the data, leading to more accurate and reliable forecasts.

The study highlighted the superiority of LSTMs in producing more accurate and reliable forecasts compared to other models such as SVR, RF, and XGBoost, particularly in scenarios involving intricate and fluctuating demand patterns.

2.5.3 ANN

Artificial Neural Networks (ANNs) are among the most widely used machine learning models for electricity demand forecasting due to their ability to model complex, non-linear relationships. ANNs are flexible and can be trained to recognise patterns in large datasets, making them suitable for applications where the underlying data distribution is unknown or highly complex. In their study, Skaf et al. (2023) demonstrated the effectiveness of ANNs in forecasting electricity consumption, particularly in capturing non-linear relationships between variables such as temperature, population growth, and electricity demand.

2.5.4 SVR

SVR is a variation of the Support Vector Machine (SVM) designed specifically for regression tasks. It is effective in capturing non-linear relationships by transforming the data into a higher-dimensional space. SVR works by finding the hyperplane that best fits the data, minimizing the error within a specified margin. This makes it particularly effective in scenarios where the relationship between input variables and electricity demand is not strictly linear. Liu et al. (2022) applied SVR in Jiangsu Province to forecast electricity consumption, emphasizing its ability to handle non-linear relationships and deliver accurate predictions in complex environments.

2.5.5 GBR

Gradient Boosting constructs an additive model incrementally, combining weak learners (usually decision trees) in a step-by-step process. It reduces errors by successively adding trees that correct the errors of previous trees. Di Persio and Fraccarolo (2023) utilised GBR to forecast energy consumption in Italy, demonstrating its effectiveness in improving forecast accuracy by focusing on difficult-to-predict data points.

2.5.6 LightGBM

LightGBM uses a tree-based learning algorithm designed for efficiency and scalability. It employs a leaf-wise growth strategy (Géron, 2022), which can converge faster than level-wise growth. It is designed for efficiency and scalability, making it well-suited for forecasting tasks that involve large datasets and high-dimensional features. Studies have highlighted the strengths of LightGBM in forecasting electricity

consumption in various regions, noting its ability to handle large-scale data with minimal computational overhead while maintaining high accuracy (Muaz et al., 2024; Skaf et al., 2023).

In the surveyed literature, Random Forest emerged as the most commonly applied machine learning model for electricity demand forecasting, highlighting its effectiveness in capturing complex, non-linear patterns within the data. Its popularity can be attributed to its robustness, ability to handle high-dimensional datasets, and ease of interpretation. LSTM also shows significant usage, indicating its strong ability to handle sequential data, particularly in time series forecasting. Other models such as ANN, KNN, and Decision Tree, while less commonly applied, are chosen for specific contexts where their unique strengths align with the forecasting requirements. This variety of models reflects the adaptability and innovation within the field, as researchers continue to explore the most effective techniques for accurate electricity demand prediction.

2.6 HYBRID MODELS FOR FORECASTING ELECTRICITY DEMAND

Hybrid models for electricity demand forecasting combine the strengths of various models or algorithms to improve prediction accuracy by addressing different data characteristics. These models may involve ML-to-ML combinations, statistical-to-statistical or ML-to-statistical hybridizations. By blending different approaches, such models capture both linear and nonlinear patterns in the data more effectively than standalone models. The flexibility of hybrid models makes them better suited to handle diverse factors, leading to more accurate and reliable forecasts for complex systems like energy demand.

There are several methods to achieve hybrid combinations of different models/algorithms, including:

Ensemble methods: This is one of the common hybridizing techniques, which involve training multiple models and aggregating their outputs to form a final prediction. This aggregation can be performed using averaging or weighted voting mechanisms. Bagging and boosting methods, like Random Forest and Gradient Boosting, are typical examples of ensemble methods that reduce model variance and improve overall performance. Studies like Ahmed et al. (2020) have shown the success of ensemble methods in energy forecasting by capturing variability in demand. Similarly, Giamarelos et al. (2023) utilised a dynamic weighting system to combine machine learning models, such as neural networks and random forests, improving accuracy across various forecasting horizons.

Model blending: This approach involves using the outputs of several models and averaging them or using a weighted average to generate final predictions, enhancing prediction accuracy by considering multiple model perspectives. For instance, a study blending linear regression with LSTM for electricity demand

forecasting improved the accuracy by capturing residual demand fluctuations after removing seasonality and trends (Hyndman & Athanasopoulos, 2018).

Stacking: In this approach, the outputs from one or more base models are used as inputs for a higher-level or meta learning model. This allows for a multi-layered system, where each subsequent model aims to correct the errors of the previous ones. The study by Guo et al. (2024) utilised a stacking method in their hybrid model for short-term load forecasting. They combined several machine learning models, including XGBoost, LSTM, Bi-LSTM, and stacked LSTM as base learners. These models were selected for their complementary strengths - XGBoost for handling structured data and nonlinear relationships, and the various LSTM models for capturing long-term dependencies in sequential data. The final predictions from these models were then combined using Lasso regression as the meta-learner. By integrating these diverse model types through stacking, the method enhanced predictive accuracy and was particularly effective at accounting for external factors like temperature, electricity prices, and rainfall, leading to a significant reduction in forecast errors.

Sequential Hybridisation: This method involves feeding the output from one model as input into another model. A common combination is using statistical models (like ARIMA) to capture the linear trends in the data, and then the residuals are passed to a machine learning model such as Random Forest or XGBoost to capture the nonlinear relationships. Yıldız Doğan et al. (2024) applied a CNN model to capture spatial patterns before passing outputs to LSTM to account for temporal dependencies. This method significantly improved predictive accuracy for electricity demand in sustainable cities by capturing both spatial and temporal features.

2.7 CURRENT FORECASTING METHODS USED IN NAMIBIA AND RELATED COUNTRIES

Electricity demand forecasting plays a critical role in ensuring grid stability, energy planning, and efficient resource allocation. In Namibia, NamPower primarily relies on historical consumption trends, load profiling, and statistical models such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) to predict electricity demand (Nampower, 2020). However, as electricity consumption patterns evolve, there is a growing need for more advanced hybrid forecasting models that integrate both statistical and machine learning techniques.

In South Africa, daily peak electrical load forecasting has been conducted using SARIMA, SARIMA–GARCH, and regression-SARIMA - GARCH models, which incorporate time-series trends and volatility adjustments for more precise forecasting. Similarly, in Botswana, forecasting methods include time series

analysis (such as ARIMA), regression models, econometric modeling, and bottom-up approaches that account for population growth, economic development, and weather patterns.

These diverse approaches across Southern Africa highlight the evolution of electricity forecasting methods, from traditional statistical techniques to more integrated, data-driven solutions. While Namibia currently leans towards statistical models, incorporating hybrid ML techniques - as seen in neighbouring countries could enhance forecasting accuracy, improve grid reliability, and support sustainable energy planning.

2.8 ANALYSIS OF KEY FACTORS INFLUENCING ELECTRICITY DEMAND

An overview by Khatoon et al. (2020) provides an extensive review of the different factors influencing electricity demand forecasting. The study highlights that meteorological factors such as temperature, humidity, and wind speed play a significant role in short-term load forecasting, while economic factors like GDP and industrial development are more influential in long-term forecasting. The paper also discusses the impact of random factors such as cultural events and special occasions on load forecasting accuracy.

2.8.1 Weather conditions

Weather conditions such as temperature, humidity, and precipitation, have a significant influence on electricity demand patterns. Several studies highlight the critical role of these meteorological factors in shaping consumption trends especially in the short-term forecasting horizon. For instance, research conducted in Jordan identified strong correlations between electricity consumption and various temporal factors, including time of day, day of the week, and seasonal variations, along with a notable relationship with daily weather conditions. Similarly, a study in Sri Lanka integrated weather variables into machine learning models, revealing that the Vector Auto Regression (VAR) model, which incorporated these patterns, outperformed other models in short-term forecasting. In Tokyo, Japan, another study quantified the impact of temperature, humidity, and precipitation on electricity demand, emphasising the significant influence of extreme temperatures on daily maximum demand. These studies collectively underscore the importance of incorporating weather conditions into short-term electricity demand forecasting models to enhance accuracy and reliability (Abeywickrama & Dinesh Asanka, 2023; Banimustafa & Al-Omari, 2022; Jie et al., 2023).

2.8.2 Economic and demographic shifts

Economic indicators such as GDP, per capita income, and population growth, are key factors influencing electricity demand. Multiple studies have demonstrated that as GDP and population increase, the demand

for electricity rises due to higher economic activity, industrial expansion, and greater energy needs in residential and commercial sectors. Studies have consistently demonstrated the strong correlation between these variables and electricity consumption, making them crucial predictors in forecasting models. For instance, models like MLP and (ANN) have been effectively employed to predict electricity consumption by integrating these economic and demographic variables. This approach is particularly valuable in emerging economies where rapid economic growth and population increases drive substantial changes in energy demand patterns. The integration of such variables ensures that forecasting models can provide more accurate and reliable predictions, thereby aiding in better energy planning and policy-making (Energy Agency, 2024; Hasan et al., 2021; Jaisumroum & Teeravaraprug, 2017).

2.8.3 Renewable energy integration

Renewable energy sources have become essential for sustainable energy systems, providing environmental advantages and lowering greenhouse gas emissions (Petrović-Ranđelović et al., 2020). However, integrating renewable energy sources like wind and solar introduces complexities into electricity demand forecasting. The variability and intermittency of these sources can create uncertainties in supply, complicating the accurate prediction of electricity demand. According to Iweh et al. (2021), "Even though the integration of distributed generation (DG) into the power grid has benefits, such as line loss reduction, improved grid flexibility, cost-effective energy production and minimization of the tendency to invest in additional grid assets during capacity expansion, several issues such as compromised quality of supply due to variations in voltages and frequencies arise when these systems are added to the network."

To investigate the effect of integrating Renewable Energy on electricity demand forecasting in an integrated smart grid Lanka et al. (2021) utilised LSTM networks to forecast solar and wind power outputs, as well as electricity load demand, to optimise the operation of the grid. The research emphasises the importance of accurate renewable energy forecasting for maintaining grid stability and minimizing operational costs. The proposed model demonstrated superior performance with higher R-squared values and lower Mean Absolute Percentage Errors (MAPE) compared to traditional forecasting methods, particularly in predicting solar and wind power generation.

2.9 RELATED WORK AND GAPS IN RELATED WORK

The works most related to this dissertation, in terms of the research objectives, include studies that combine traditional statistical models with machine learning techniques to improve the accuracy of electricity demand forecasting. One notable study conducted by Krishnan et al. (2020) focused on combining

SARIMA and LSTM models, to create an advanced hybrid SARIMA-LSTM model for time-series data forecasting. This model leverages SARIMA's strength in handling linear and seasonal components, while LSTM captures nonlinear patterns. The study demonstrated that the hybrid model significantly outperformed the standalone SARIMA and LSTM models in terms of RMSE, MSE, and MAE, indicating its effectiveness in improving prediction accuracy for energy demand in smart grids.

Similarly, Hou et al. (2023) proposed a hybrid ARIMA-LSTM (ARLS) model for urban electricity demand forecasting, combining traditional time-series methods with machine learning techniques. The ARIMA model was used to capture the trend, seasonal, and level components, while the LSTM network handled the nonlinear features. By combining the outputs using an ensemble technique, the ARLS model achieved high prediction accuracy, making it a robust tool for long-term urban electricity demand forecasting.

Despite these advancements, a significant gap remains in the literature, particularly in applying these hybrid models to the unique context of Namibia. Related studies have successfully demonstrated the benefits of combining traditional statistical models with machine learning techniques, they often limit their scope by focusing on the electricity consumption as an input variable to the models alone, without fully exploring the potential of incorporating additional contextual variables such as economic indicators, population growth, and renewable energy integration.

This dissertation builds on the foundational work of these studies by extending the hybrid approach to include additional factors, such as economic indicators, population growth, and renewable energy integration, alongside historical electricity consumption data that are particularly relevant to the City of Windhoek, Namibia. This integration creates a more comprehensive and accurate tool for electricity consumption forecasting. The study also systematically identifies the best-performing models from both traditional and machine learning categories to develop an optimised hybrid model, specifically tailored to the Namibian context.

Moreover, the financial implications of managing NMD in the City of Windhoek underscore the need for more accurate and region-specific forecasting models. By addressing these unique challenges, this study contributes to the broader understanding of electricity demand forecasting and supports more effective energy management practices, filling the gap in the existing literature.

2.10 CHAPTER SUMMARY

Electricity demand forecasting is a crucial area of research, as evidenced by numerous studies that address the challenges and methodologies involved. The approaches to electricity demand forecasting can be broken down into three generic categories namely: traditional statistical models, ML-based models, and hybrid models.

Traditional statistical models such as ARIMA, SARIMA, and HWES, are widely used due to their ability to handle linear time series data and seasonal patterns. Machine learning models, such as Random Forest, XGBoost, and LSTM, offer advanced capabilities to capture complex, and non-linear relationships in the data. Hybrid models combine the strengths of both traditional and ML-based approaches, aiming to enhance forecast accuracy by leveraging the advantages of each.

Despite the significant progress made in this field, there remains a research gap in the application of hybrid forecasting models to specific contexts like Namibia, where unique economic, infrastructural, and energy-related factors necessitate further exploration. The present study aimed to address these gaps by optimising and applying hybrid models for electricity demand forecasting tailored to the specific needs of the City of Windhoek.

CHAPTER THREE: METHODOLOGY

3.1 CHAPTER OVERVIEW

This chapter outlines the methodological framework used to address the research objectives outlined in the preceding section. The methodology combines quantitative and qualitative approaches to provide a comprehensive understanding of the complex dynamics influencing energy consumption patterns for the City of Windhoek, with specific emphasis on forecasting electricity demand. The study combines the principles of Saunders' (2009) research onion with the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework (Shearer, 2000) to develop a comprehensive and structured research methodology for forecasting electricity demand in Windhoek. While the research onion provides the philosophical foundation, CRISP-DM offers a detailed roadmap for handling the data-driven aspects of the study. The map shown in Figure 3.1and Figure 3.2 below outlines the chapter layout and the research process respectively.

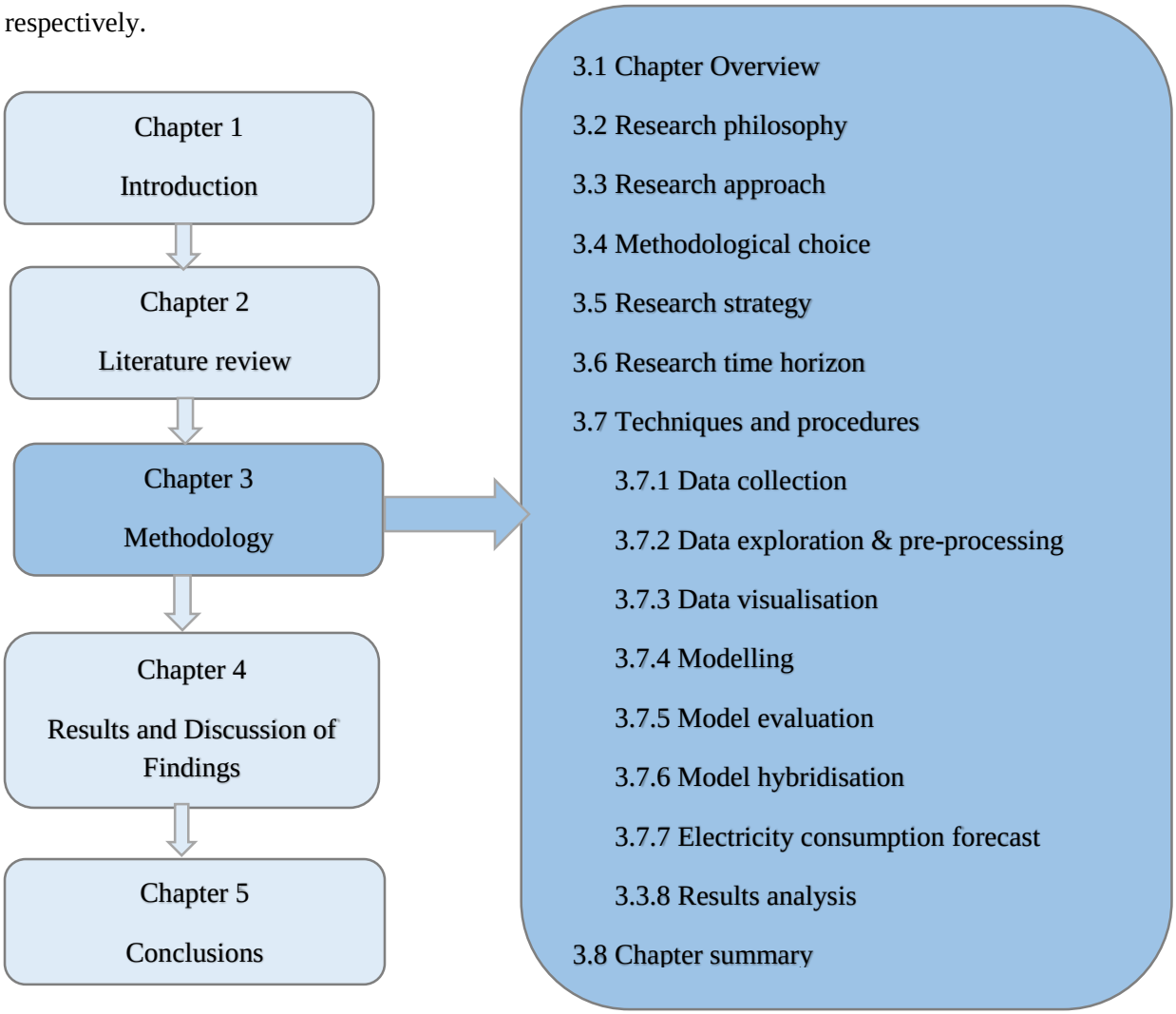


Figure 3. 1: Chapter 3 outline map. Source: author

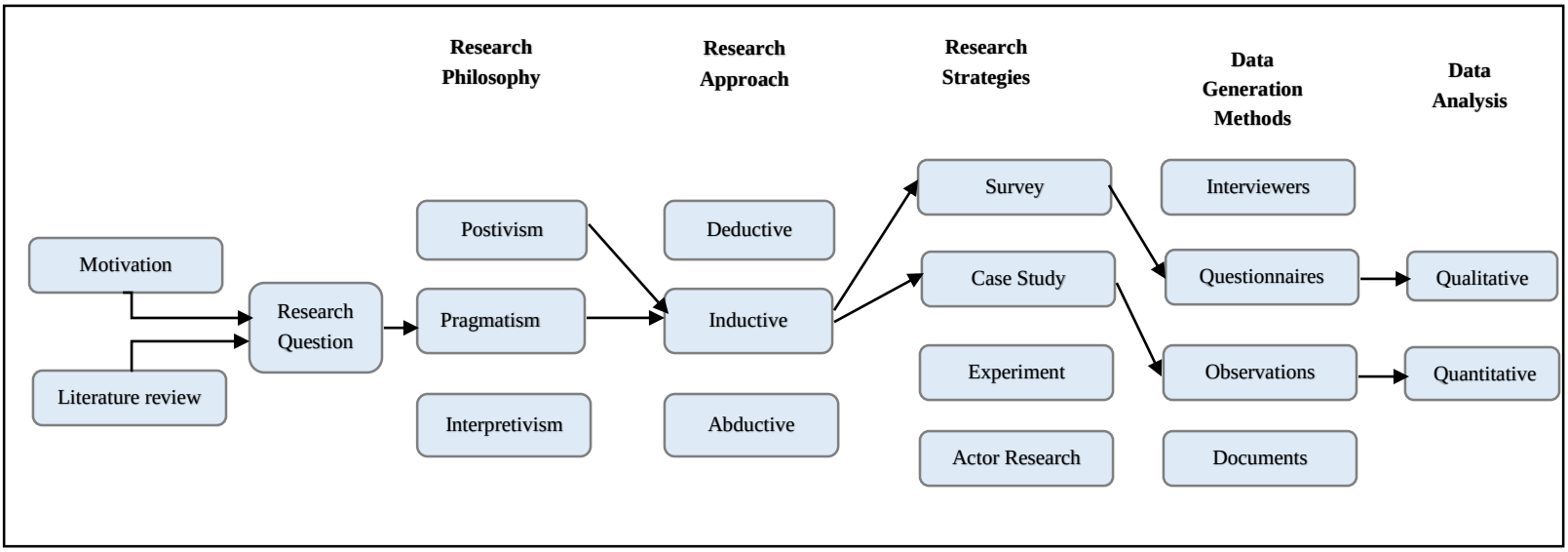


Figure 3. 2: Model of the research process. Source: Saunders et al., 2019)

3.2 RESEARCH PHILOSOPHY

Research philosophy serves as a guiding principle for researchers, aiding in selecting the most appropriate approach and techniques for their studies. According to Saunders (2009), a research philosophy refers to a framework of beliefs and assumptions about how knowledge is developed. This study adopts the principles of both pragmatism and positivism. The pragmatist philosophy, focusing on practical outcomes, guided the investigation into effective solutions and applications for managing electricity demand in the City of Windhoek. It allowed for flexibility in considering multiple approaches to problem-solving, ensuring that the research remained focused on producing actionable results. Concurrently, positivism supported the objective analysis of historical electricity demand data and economic factors, enabling a systematic and empirical approach to generating meaningful insights.

3.3 RESEARCH APPROACH

The process of theory development typically encompasses three main approaches: deduction, induction, and abduction. According to Melnikovas (2018), in the deduction approach, researchers begin with an established theory, formulate a hypothesis or question, and then collect data to either confirm or reject the hypothesis. Conversely, induction involves starting with observations and data collection, progressing to description and analysis, and forming a theory based on the findings. Abduction, meanwhile, entails observing empirical phenomena and then conducting research that leads to conclusions based on the

available facts. These approaches offer different paths for researchers to develop and refine theories, each with its own strengths and applications.

The research approach for this study primarily follows an inductive methodology, which entails commencing with observations and data collection, description and analysis, and subsequently deriving a theory based on the discovered patterns and insights. Given the limited existing theory specifically for the City of Windhoek's context, the complexity of factors influencing electricity demand, and the pragmatic philosophy guiding the research, an inductive approach allows for the systematic exploration of real-world observations and data analysis to derive practical insights. By focusing on the City of Windhoek's distinctive features and dynamics, this approach enables the development of contextually relevant theories and models that address the city's specific energy forecasting requirements.

3.4 METHODOLOGICAL CHOICE

In accordance with the pragmatism and positivism research philosophies, this study employed a Mixed Methods Approach. This approach involves the collection, analysis, and interpretation of both quantitative and qualitative data across multiple studies focused on the same fundamental phenomenon (Clark & Clark, 2022). Integrating both qualitative and quantitative methods provides a comprehensive understanding of the various factors influencing electricity demand in Windhoek. By utilising a Mixed Methods Approach, the research leveraged the strengths of both data types to gain insights into the complex interactions between economic, seasonal, and demographic factors affecting electricity consumption.

3.5 RESEARCH STRATEGY

The study employed three strategies, namely Case Study, Survey Analysis, and Archival Research.

Case study: This involved analysing historical electricity demand data for the City of Windhoek over the past decade to identify specific trends and patterns. The case study approach allowed for an in-depth examination of a single case, providing extensive information and insights (Organization for Autism Research, 2020).

Survey analysis: Surveys were conducted with key energy experts to gather their views on electricity demand, economic growth, and the integration of renewable energy in Windhoek.

Archival research: This involved a systematic investigation of primary sources in archives to collect evidence and gain a deeper understanding. This involved analysing historical data on electricity consumption, economic indicators and population growth.

The research follows a Mixed Methods Approach, integrating both qualitative and quantitative techniques for data collection and analysis. Qualitative data from surveys provided valuable insights into the perspectives and experiences of energy experts regarding electricity consumption patterns and influencing factors. On the other hand, quantitative data from historical records, economic indicators, and renewable energy integration were analysed using statistical and machine learning methods to uncover patterns, correlations, and trends, enhancing the accuracy of electricity demand forecasting.

3.6 RESEARCH TIME HORIZON

The Time Horizon layer of the research onion as outlined by Melnikovas (2018), is crucial for defining the scope and establishes the timeframe for the research. It delineates between two primary time horizons: cross-sectional and longitudinal.

This study employed a cross-sectional time horizon, which is defined as data collected at a specific moment in time, without subsequent follow-up (Saunders, 2009), thus aligning with the research's focus on utilising existing data from the City of Windhoek database at a single point. By opting for a cross-sectional time horizon, the study aimed to capture a snapshot of electricity consumption patterns and associated factors within Windhoek's urban landscape, and to obtain insights that can be incorporated in the forecasting of energy demand.

3.7 TECHNIQUES AND PROCEDURES

The techniques and procedures, which is the final layer of the research onion, constitute the techniques and procedures that guides the research design towards data collection and analysis. This section outlines the systematic approach aimed at realising the study's objectives. This layer encompasses the sequence of research events and their execution to achieve the research goals. The following key components of Techniques and procedures were employed in this study.

3.7.1 Data collection

According to Muhammad and Kabir (2018), data collection involves the logical gathering and measurement of information pertaining to variables of interest in a systematic manner. This process enables researchers to address specified research questions and assess outcomes. The process typically commences with determining the type of data needed, followed by the selection of a sample from a target population.

Following the protocols outlined by the Faculty of Research Ethics Committee (F-REC) of NUST, the data collection process typically commenced with the submission of an ethical clearance application to the F-REC. All the necessary documents required for the ethical clearance review, including the research proposal, informed consent forms, and survey questionnaires, were compiled and submitted. The researcher proceeded with the next step of collecting the data after the ethical clearance (Appendix A) was issued.

The primary methods employed in collecting data for this study included gathering existing historical energy consumption data from the City of Windhoek spanning the past decade, surveying the energy experts, and sourcing economic indicators and population growth data online. The participants for the survey were required to read the Information consent form where the researcher had fully explained the purpose of the study, their rights as participant and the confidentiality measures put in place to safeguard their identity. These methods were chosen to ensure a comprehensive dataset that reflects both the historical trends in energy consumption and the contemporary factors influencing electricity demand in order to achieve the study's objectives. These methods and data sets used are explained in detail below:

(a) Historical electricity consumption: This process involved accessing and compiling records of electricity consumption spanning the past decade (2013-2024) in monthly intervals. Electronic data was only available for the past four years; however, to complete the dataset, the remainder of the records were retrieved from archived hard copies. The data was extracted from the Nampower Invoices sent to the City of Windhoek, which served as the primary source of historical consumption records. These records comprised essential variables necessary for the analysis of electricity trends and demand forecasting as summarised in Table 3. 1.

Table 3. 1: CoW historical electricity consumption dataset selected features

Feature	Description
Billing date	The month and year of the billing period
Season:	The season for the billing month classified as either high or low. High season corresponds to higher usage and rates due to extreme weather, while low season sees lower demand and rates during milder conditions.
Main Meter Readings values in kWh	The main energy meter installed on Nampower' side readings.
Check Meter Readings values in kWh	The check meter installed on City of Windhoek side readings, for confirmation of Nampower's billing.
Max. Demand Date and Time:	The time and the date at which the actual maximum demand for the specific month occurred.
Power factor @Max. Demand:	The power factor defined as the ratio of real power consumed by the load to the apparent power flowing in the circuit that was experienced at Max. Demand
Demand Charge:	The cost incurred by Nampower to meet City of Windhoek's peak electricity demand calculated based on the contracted capacity.
Energy Charge in kWh, N$/kWh and N$ at Peak, Standard and Off Peak:	The amount of energy consumed multiplied by the applicable tariff, which varies depending on the day (peak, standard, and off-peak rates). The sum of this energies gives the total energy consumed in the month.
ECB Levies	Regulatory charges imposed by the Electricity Control Board
NEF Levies	Contributions towards the National Energy Fund, which supports initiatives aimed at supporting sustainable energy development and energy efficiency projects at the national level.
Total for the Month	The total amount for the month that should be paid to Nampower which includes all charges incurred during the billing month.

Renewable Energy	The amount of renewable energy installed in the City of Windhoek electrical grid. This data was only available on the yearly basis so the researcher interpolated the yearly values to match the monthly dataset.

(b) Economic indicators and population: Table 3. 2 shows the features of the historical data on economic indicators and population that was incorporated in the analysis. The data was collected from macrotrends website (Macrotrends, 2024) and included variables that were deemed significant for analysis purposes. It is important to note that while this study focuses on the City of Windhoek, the economic indicators and population data are at the national level. This approach assumes that national trends can serve as a proxy for patterns at the city level. However, it is acknowledged that population and economic growth dynamics can differ between the national and city levels. Therefore, the results derived from national-level data are interpreted as indicative of city-level trends.

Table 3. 2: Economic indicators and population dataset features

Feature	**Description**
Date	The specific year to which the population and economic indicators correspond.
Population	The total number of inhabitants in Namibia
Population Annual Growth Rate	This metric indicates the percentage increase or decrease in the population size over a one-year period. It reflects the rate of population change relative to the total population size.
GDP (Gross Domestic Product) in Billions of US $:	The total monetary value of all goods and services produced in Namibia annually. It is measured in billions of US dollars and serves as a key indicator of a country's economic performance and productivity.
Per Capita (US $):	The average income per person in the country.

Annual % Change	Percentage change in GDP or per capita compared to the previous year.

The economic indicators and population data was also only available on the yearly basis, so the researcher interpolated the yearly values to match the monthly energy consumption dataset for analysis and models training.

(c) Survey: The study employed an online survey which was setup using Microsoft forms and distributed to carefully selected participants. The survey targeted 40 participants chosen based on their expertise and involvement in electricity demand management, economic growth, and renewable energy integration in Windhoek. The insights provided by these experts were integral to the study's analysis, offering valuable perspectives on critical issues. Responses were collected via Microsoft Forms and then exported to a CSV file for analysis. The survey questions are included in this document as Appendix B. Figure 3. 3 below summarises the data collection process and the data sources used.

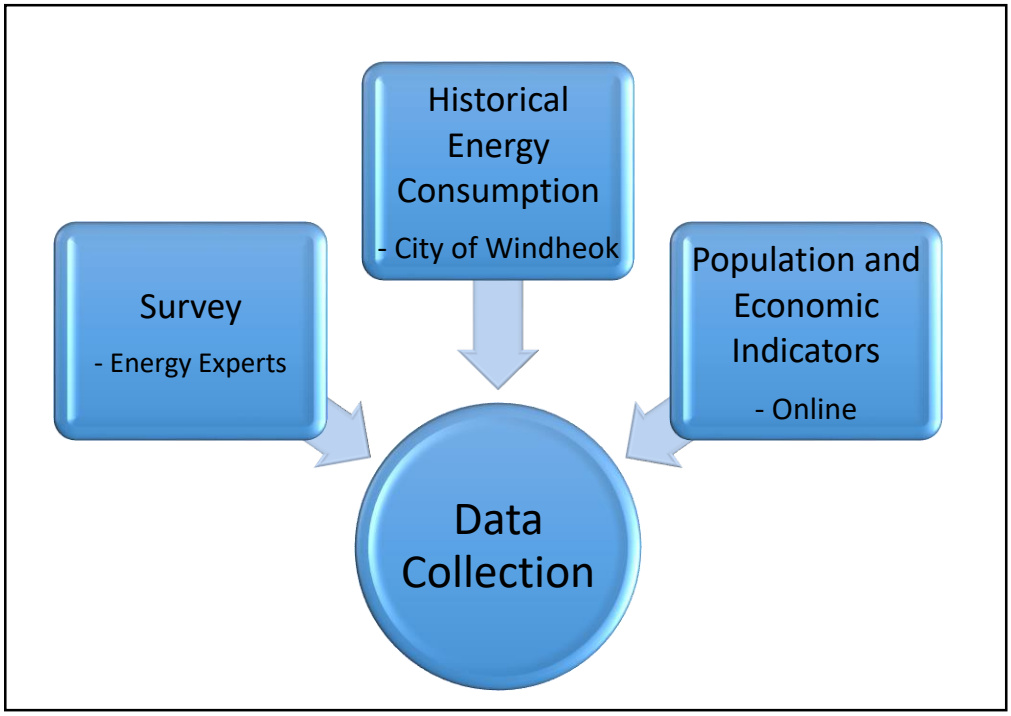


Figure 3. 3: Data collected and sources. Source: author

3.7.2 Data cleaning and pre-processing:

Data exploration and pre-processing are crucial steps in the research process, aimed at understanding the characteristics of the dataset, identifying patterns, and preparing the data for further analysis. The data pre-processing was carried out using python along with relevant libraries. Following a comprehensive data exploration, various issues were identified and handled as summarised in Table 3. 3 and Table 3. 4. The steps taken ensured that the dataset was accurate, complete, and suitable for use in the forecasting model.

One of the key pre-processing steps involved data normalisation, which is critical for ensuring that the features contribute equally to the analysis and model training. Specifically, **Min-Max Normalization** was used in this study to scale the numerical variables to a range between 0 and 1. The Min-Max Normalization method is formulated in Equation 3.1 below:

$$X' = \frac{X - X_min}{X_\max - X_min} \qquad \text{Equation 3.1}$$

Where X' is the normalised value,

X is the original value,

X_min is the minimum value

X_max is the maximum value

The method was implemented using MinMaxScaler class from the scikit-learn library. Normalisation using this method ensures that all features have the same scale, which is particularly important for models sensitive to feature magnitudes, such as those using distance-based metrics or gradient descent. Table 3.3 shows a summary of the data quality issues encountered while pre-processing the CoW raw data on historical energy consumptions. Table 3. 3 is to be read in conjunction with Appendix C that gives full details and explanation of the cleaning action performed on the CoW historical energy consumption dataset.

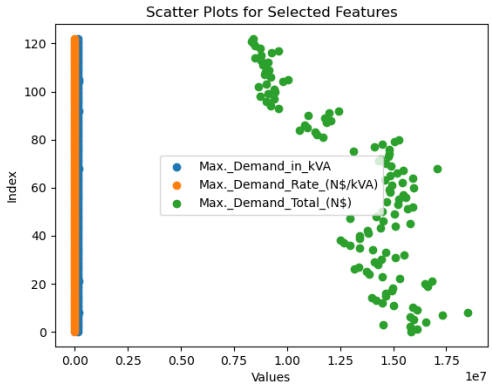
Table 3. 3: Pre-processing of CoW historical energy consumption raw data

Cleaning Action Performed	**Description**
Renaming columns	Column names were standardised by removing problematic characters and spaces to ensure consistency and compatibility with various data processing libraries.
Combining Relevant Columns due to tariff unbundling	In 2020, the City of Windhoek undertook a tariff unbundling exercise, which resulted in the separation of electricity tariffs into distinct components to enhance transparency and better reflect the actual cost structure of electricity supply. The tariffs were unbundled into two main categories: Energy and Losses. Prior to 2020, the data values were bundled and did not separately account for the losses component; it was included in the total energy values. To ensure uniformity for modeling and to accurately reflect the post-2020 tariff unbundling into energy and losses components, these columns were added. Refer to the code snippet in Appendix D.1.
Missing values	Missing values were handled by way of imputation with numerical columns imputed using the mean value of each respective column and for the categorical columns, missing values were imputed using the mode. The observed missing values in the dataset across both numerical and categorical columns, were significantly small (refer to Appendix D.2), which means imputing them did not comprise on the quality and integrity of the data.

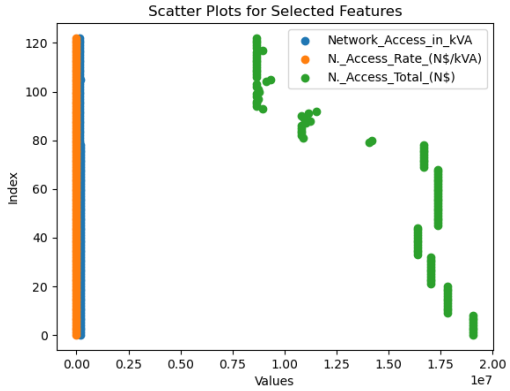
Duplicate rows	The dataset did not contain any duplicate rows.
	
	`# Check for duplicates`
	`print("\nDuplicates:")`
	`print(cow_energy_consumption.duplicated().sum())`
	
	`Duplicates:`
	`0`

Outliers	Different features in the datasets relevant to energy consumption forecasting were plotted and visualised in the following scatter plots to identify outliers. Most features did not indicate the presence of data points observed at unusual distance from the rest of the observations except for plot E and F as depicted below.

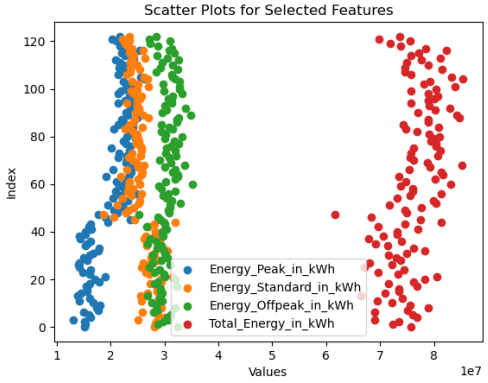
Plot A



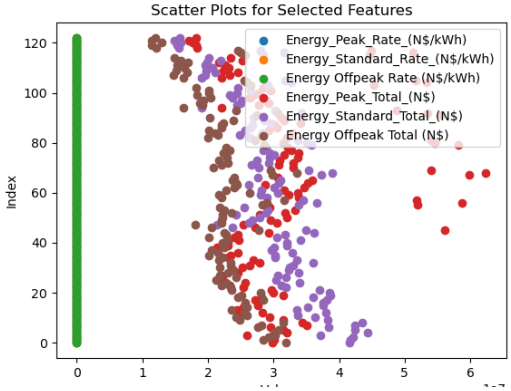
Plot B



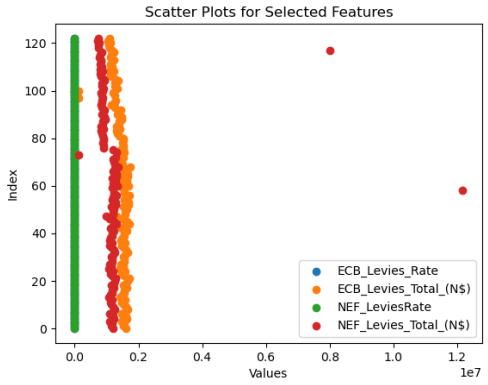
Plot C



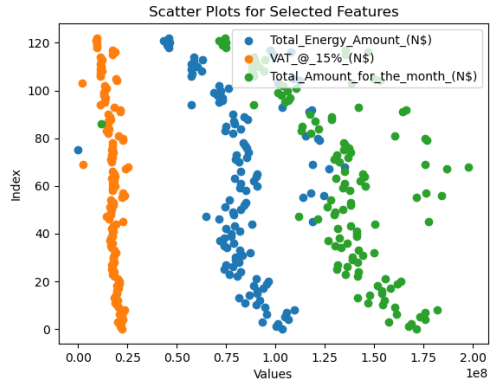
Plot D



Plot E



Plot F



The identified outliers were investigated to determine their causes, thus revealing that most discrepancies originated from manual data entry errors from the invoice records. Corrective actions were taken by referencing and rectifying data directly from the original manual invoices. For outlier observations not due to entry mistakes, they were rectified by replacing them with the mean of that feature. This ensured that erroneous entries were corrected, thus improving the overall accuracy and consistency of the dataset. Refer to Appendix D.3, for the outliers' handling code snippet.

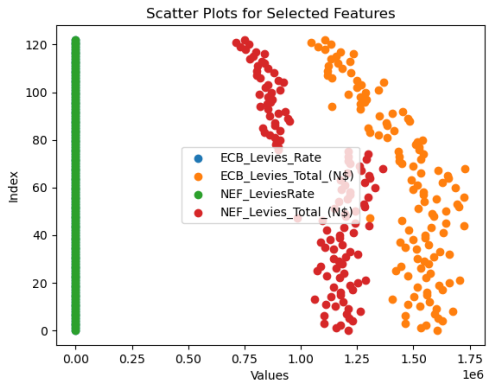
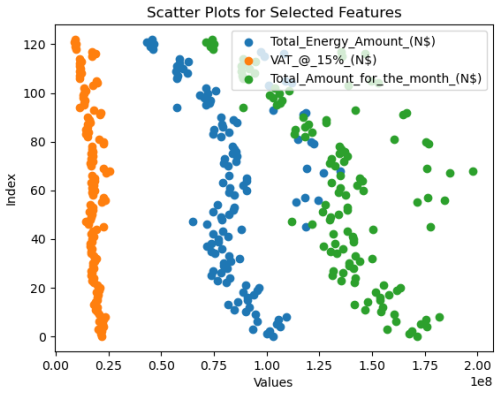
	Plot G and H below depict plot E and F after addressing the outliers, respectively. Plot G  Plot H 
Dropping irrelevant columns	Several columns deemed not meaningful or redundant were dropped to streamline the dataset, thus making it more focused and relevant for the task of energy consumption forecasting.

Table 3.4 shows a summary of the data quality issues encountered while pre-processing the Economic Indicators, Population and Renewable Energy raw data.

Table 3. 4: Pre-processing Economic, Population and Renewable Energy Raw Data

Cleaning Action Performed	**Description**
Missing values, duplicate rows and outliers.	The dataset did not contain any missing values, duplicate rows or any outliers as depicted in the below code snippet and scatter plots.

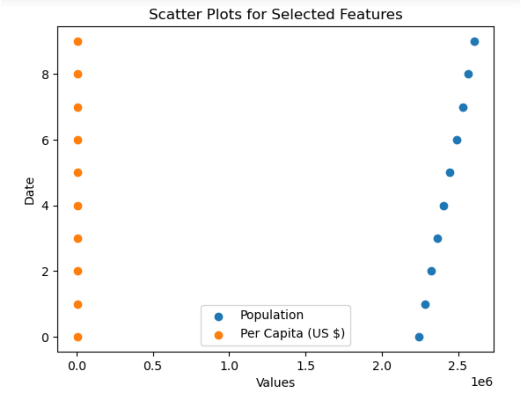
```
# Check for missing values
print("\nMissing values:")
print(nam_gdp_population.isnull().sum())
```

```
Missing values:
GDP ( Billions of US $)    0
Per Capita (US $)          0
Annual % GDP Change        0
Population                 0
Annual Pop Growth Rate     0
dtype: int64
```

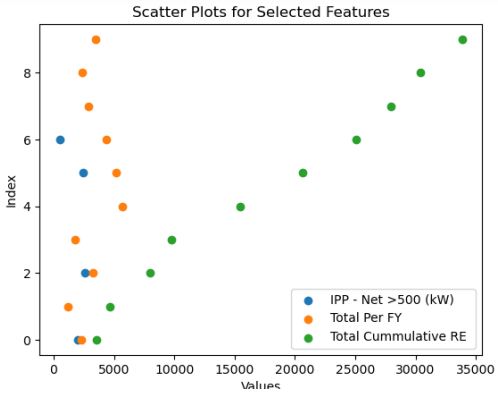
```
# Check for duplicates
print("\nDuplicates:")
print(nam_gdp_population.duplicated().sum())
```

```
Duplicates:
0
```

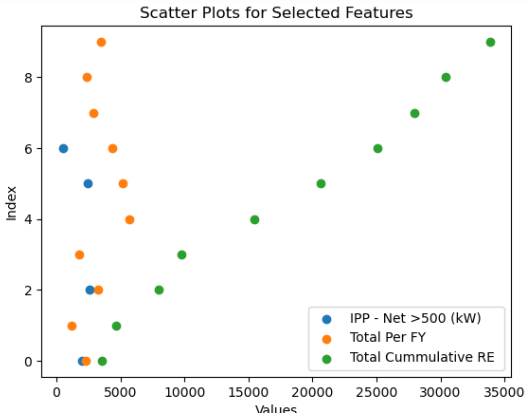
Plot I



Plot J



Plot K



Dropping irrelevant rows	The collected dataset had entry rows of GDP and population dated until the 1980s but since the study focuses only on the past ten years, rows dated on or before December 31, 2013 were dropped. The final dataset contained only the rows from 1 January 2014 onwards, relevant for the analysis.
Data Interpolation	The data for GDP, population, and renewable energy were originally collected on a yearly basis, while the energy consumption data was available on a monthly basis. To ensure compatibility of these datasets for meaningful insights, comparisons, and modeling, interpolation was applied. This involved interpolating the yearly data to a monthly frequency, allowing all datasets to be aligned on a consistent monthly timeline. Linear interpolation was chosen to convert yearly data on renewable energy, GDP, and population to a monthly frequency because it provides a smooth, gradual transition between data points, preserves the overall trend, is simple and effective to implement, and ensures compatibility with time series modeling. Refer to Appendix D.4 for the code snippet of the interpolation process.

This process of data pre-processing is summarised in the flowchart in Figure 3. 4.

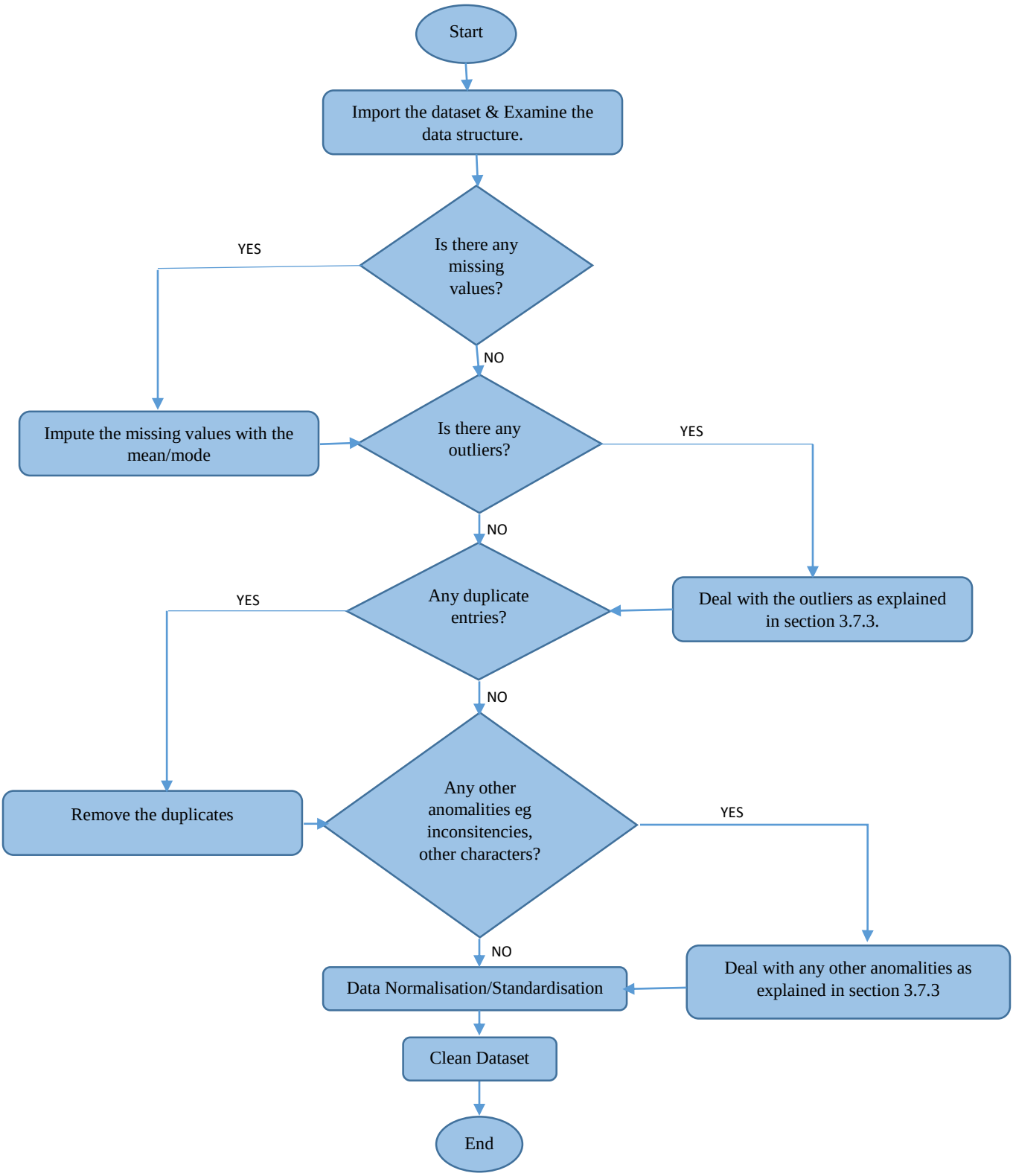


Figure 3. 4: Data pre-processing flowchart. Source: author

3.7.3 Data visualisation:

Data visualisation and analysis is a crucial step in the research process, it helps the researcher gain a deeper understanding and insight from the data.

According to Wang et al. (2016), data visualisations entails presenting data in graphical formats like charts and graphs to simplify complex data structures and relationships. Its purpose is to enhance comprehension across various fields by leveraging our visual capabilities to quickly identify colors, shapes, and patterns swiftly. Through data visualisation, information becomes more accessible and facilitates improved exploration and interpretation of data.

In this study, visualisation was implemented in Python, using libraries for data visualisation and plotting such as Matplotlib, Seaborn, Plotly, and pandas. These libraries offer various capabilities and cater to different preferences and requirements for data visualisation in Python. The visualisation is done using different plots such as histograms, box plots, scatter plots, and heatmaps to identify patterns, trends, and correlations.

Further visualisation was employed using the pandas library in Python, which provided various functions to calculate summary statistics such as mean, median, standard deviation, variance, and more for numerical variables. By computing summary statistics for numerical variables in this study, the researcher gained valuable insights into the central tendency and dispersion of energy consumption patterns and other variables, enabling better understanding and forecasting of electricity demand dynamics in the City of Windhoek.

3.7.4 Modelling

The model employed by this study is the combination of Machine learning models and the traditional statistical models to create an advance hybrid model to forecast the energy consumption for the City of Windhoek.

Statistical models are effective for capturing the seasonal and trend components of the time series data, which is an evident factor in electricity consumption. Machine Learning, on the other hand, is powerful in capturing long-term dependencies and nonlinear patterns (Charfeddine et al., 2023). By integrating these two techniques, the hybrid model leverages the strengths of the best performing model from each category for more accurate and robust forecasting of energy demand.

The input to both the Statistical and machine learning models were the City of Windhoek historical electricity consumption data with features from other data sources e.g. Renewable energy, population and

GDP. Following CRISP-DM's modeling phase, machine learning and statistical models were trained, validated, and refined to improve accuracy in forecasting electricity demand (Shearer, 2000).

(a) Data loading and preprocessing

The datasets were loaded and the 'Date' column is converted to a datetime format, which is crucial for time series analysis. Subsequently, the 'Date' column is set as the DataFrame's index to facilitate time-based indexing and operations.

To ensure the data is on a comparable scale, the MinMaxScaler from scikit-learn was used to normalise all relevant columns to a range between 0 and 1. The normalised data is then converted back into a DataFrame, maintaining the original column names and index for ease of further processing. This preprocessing step ensured that the dataset is ready for creating sequences and training the various machine learning models implemented. Refer to Appendix D.5 for the data loading and preprocessing code snippet.

(b) Data splitting

According to Encode (2023), splitting the data into training and testing is a crucial practice in machine learning and forecasting tasks as it allows for model performance evaluation. By evaluating the model on this unseen data, it is assessed on its generalisation performance and how well it will perform on new, unseen data in the future.

The dataset was split into training and testing sets, with 85% of the data allocated for training and the remaining 15% for testing. The data split ratio is due to the small dataset size, which consists of only 123 monthly observations. This ratio was chosen because most models explored performed better with more training data to learn from.

(c) Data sequence creation for ML

To capture the dependencies over time in electricity consumption data, this study formulated the forecasting task as a time-series problem. A critical part of this formulation is determining the appropriate time window, also known as the time lag. This window defines the number of preceding time periods (months) that are used to predict future values (Brownlee, 2020). For this study, a 12-month window was selected based on observed seasonality in the data, allowing the models to learn both long-term trends and seasonal fluctuations. The sequences of 12 months were transformed into input-output pairs, where each input sequence corresponds to the previous 12 months, and the output is the consumption for the following month.

To prepare the dataset for training and testing, the time series data was transformed into sequences of the specified length of 12 months. The *create_sequences* function converts the data into input-output pairs, where each input sequence comprises 12 months of historical data, and the output is the consumption value immediately following the sequence. Further sequencing was done to ensure the data is specifically compatible with the LSTM model requirements. The *create_sequences_lstm* function was utilised to convert the normalised data into sequences of a specified length (12 months). Each input sequence comprised 12 months of historical data, and the output was the consumption value immediately following the sequence. This function ensured that the input sequences were in the appropriate 3D shape required by LSTM layers.

(d) Data stationarity for statistical models

Stationarity is a key assumption for ARIMA and SARIMA models. To check for the stationarity of the data, the Augmented Dickey-Fuller (ADF) test was performed on the 'Consumption' column. The ADF test provided statistics to determine if the time series is stationary. The code snippet in Figure 3. 5 below depicts the results for the ADF test.

The ADF statistic was found to be -0.4684. This value is used to compare against the critical values to determine stationarity (Prabhakaran, 2023). If it is less than the threshold values (critical values) at different confidence levels (1%, 5%, and 10%), indicates stationarity. Since the ADF statistic (-0.4684) is greater than all the critical values at all levels, which implies that the series is non-stationary.

```
# Step 2: Check stationarity of the original data
result = adfuller(df['Consumption'])
print('ADF Statistic:', result[0])
print('p-value:', result[1])
print('Critical Values:')
for key, value in result[4].items():
    print(f'   {key}: {value}')
```

```
ADF Statistic: -0.4683793792777321
p-value: 0.8980407469214278
Critical Values:
   1%: -3.4912451337340342
   5%: -2.8881954545454547
   10%: -2.5809876033057852
```

Figure 3. 5: Stationarity check code snippet. Source: author

Given that the time series is non-stationary, the data was transformed through differencing to achieve stationarity. After checking the stationarity again with the differenced data the results are presented in Figure 3. 6 below. The ADF statistic is -5.5838 which is significantly lower (more negative) than the critical values, indicating that the data is now stationary. The stationarity of the time series ensures that the statistical properties such as mean, variance, and autocorrelation are consistent over time, thus making the series predictable and suitable for forecasting models.

The below code snippet depicts the results after differencing.

```
ADF Statistic (Differenced): -5.583820449035686
p-value (Differenced): 1.3773462474324157e-06
Critical Values (Differenced):
   1%: -3.4912451337340342
   5%: -2.8881954545454547
   10%: -2.5809876033057852
```

Figure 3. 6: Differenced code snippet (Source: author)

The ACF and PACF plots

The ACF (AutoCorrelation Function) plot in Figure 3. 7 shows the correlation between the time series and its lagged values, to identify the presence of autocorrelation in the time series data. Autocorrelation refers to the relationship between a a time series with its own previous or future values (Prabhakaran, 2023). It is used to determine the order of the Moving Average (MA) component in ARIMA models. Significant spikes in the ACF plot indicate the presence of autocorrelation at specific lags, suggesting the need for MA terms to account for this autocorrelation.

The PACF plot helps to identify the number of autoregressive (AR) terms required in the ARIMA model. The PACF shows the correlation of the time series with its own past values, with the influence of intermediate lags removed. Significant spikes in the PACF plot indicate the direct relationship between the time series and its lagged values, suggesting the need for AR terms to capture these direct influences.

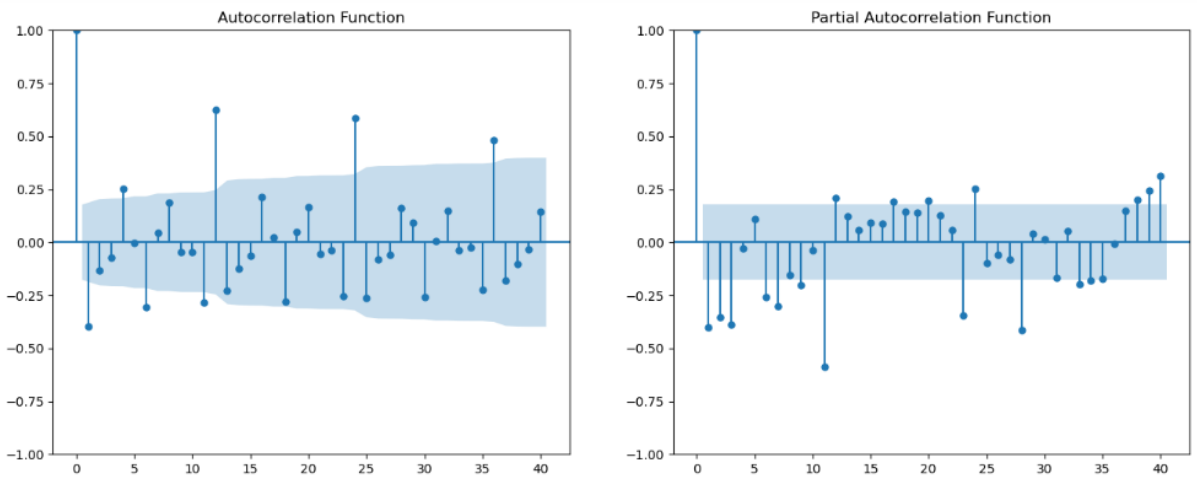


Figure 3. 7: ACF and PACF plots. Source: author

(e) Machine learning models

The following machine learning models were explored for the forecasting task of the electricity consumption for the City of Windhoek. Table 3. 5 below gives the overview of each models description, key parameters and the training process. Refer to Appendix D.6 for the ML models implementation code snippet.

Table 3. 5: ML models training methodology

Model	Key Parameters	Training Process
LSTM	- Sequence Length:12 - Hidden Units: 50 - Activation Function: (tanh)	Normalise data and create sequences. Define LSTM layers, hidden units, and additional layers. Compile the model with Adam optimiser and MSE loss. Train the model for 50 epochs.
Random Forest Regressor (RF)	Grid search with cross-validation was employed to find the optimum hyperparameters for the RF model.	1. **Hyperparameter Tuning**: Grid search with cross-validation. 2. **Model Training**: Train the Random Forest model using the best hyper parameters.
Gradient Boosting Regressor (GBR)	A grid search with cross-validation was employed to find the optimal hyper parameters for the GBR.	1. **Hyperparameter Tuning:** Grid search with cross-validation. 2. **Model Training:** Train the model with the optimal parameters identified.

LightGBM Regressor	- Number of Trees: 100 - Learning Rate: 0.05	1. **Model Training**: Train the LightGBM model with specified parameters

The graph in Figure 3. 8 shows the training and validation loss over the 50 epochs for the LSTM Model. The loss function quantifies the discrepancy between the predicted values and the actual values. From the plot, it is evident that at early training, both training and validation losses decrease rapidly, thus indicating that the model is learning well. As epochs increase, the training loss continues to decrease and stabilise, thus showing that the model parameters are converging.

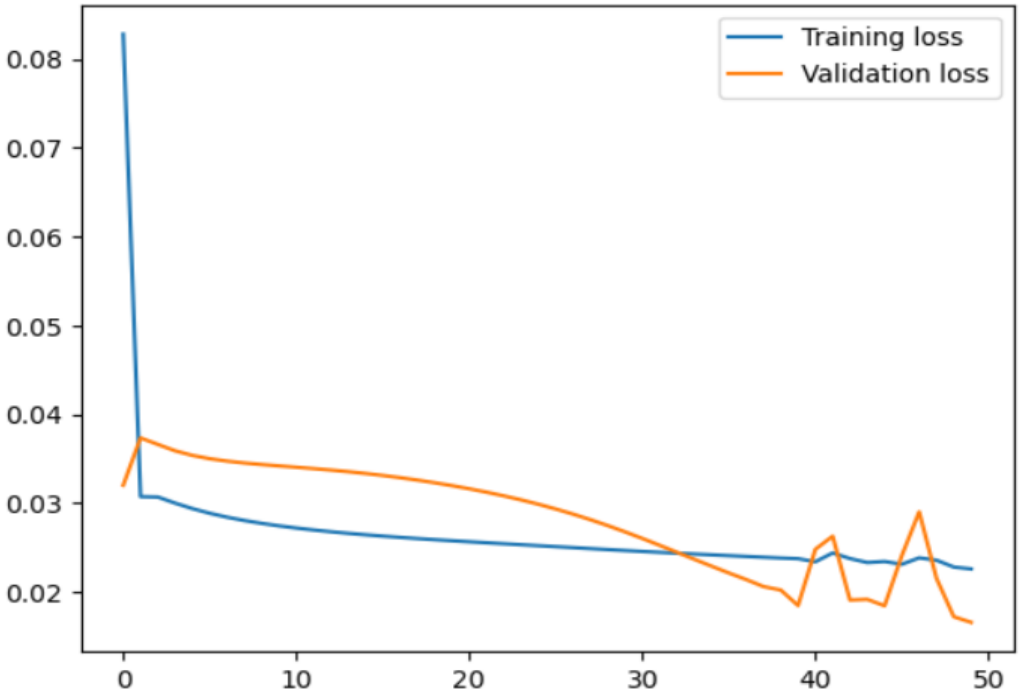


Figure 3. 8: Training and validation loss. Source: author

(f) Traditional statistical models

The following traditional statistical models were explored for the forecasting task of the electricity consumption for the City of Windhoek. Table 3. 6 below gives the overview of each model's key parameters and the training process. Refer to Appendix D.7 for the statistical models implementation code snippet.

Table 3. 6: Statistical models training methodology

Model	Key Parameters	Training Process
ARIMA	AR order (p):3 Differencing (d): 1	**1. Stationarity check:** ADF test to ensure data stationarity. **2. Differencing**: Apply differencing to achieve stationarity.

	MA order (q): 3	**3. Model identification:** Use ACF and PACF plots to identify optimal parameters. **4. Model fitting:** Fit ARIMA to training data.
SARIMA	Auto-detected using auto_arima.	**1. Auto-parameter selection**: Use auto_arima to automatically detect the best seasonal and non-seasonal parameters. **2. Model fitting:** Fit SARIMA to the training data using the identified parameters.

3.7.5 Model evaluation:

According to Alooba (2024), model performance metrics are a collection of statistical measures utilised to evaluate and compute the performance of machine learning and statistical models. These metrics assess various aspects of a model's performance and overall effectiveness. By using model performance metrics, researchers can identify the strengths and weaknesses of different models, thus allowing them to make informed decisions about which model is most suitable for a particular task or problem. These metrics offer objective comparisons that help in selecting the most reliable and accurate models for specific needs. They are designed to assess the predictions made by a data model against actual observed outcomes, thereby providing a quantitative evaluation of the model's ability to generalise and make accurate predictions on new, unseen data.

All the models were evaluated on the same evaluation metrics, MAE, MSE, RMSE, R^2 Score and MAPE calculated by the *evaluate_and_plot* function, that is crucial for assessing and comparing the performance of predictive models. The actual versus the predicted values were then plotted for all the models to visually inspect model accuracy. This function standardises the evaluation process across different models, allowing for a clear comparison of their predictive capabilities. By summarizing these metrics and generating comparative plots, it provides a broad understanding of each model's performance in forecasting electricity consumption. Refer to Appendix D.8 for the model evaluation code snippet. Table 3. 7 below summarises the evaluation metrics applied on all the explored models.

Table 3. 7: Evaluation metrics

Evaluation Metrics	**Description**	**Formula**
Mean Absolute Error (MAE)	MAE computes the average absolute difference between the forecasted and actual electricity consumption values. Lower MAE values indicate better	$$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert yi - \hat{y}i\rvert$$ Equation 3.2: Mean Absolute Error

	accuracy (Robeson & Willmott, 2023). A lower MAE means that, on average, the model's predictions are closer to the actual electricity consumption values.	Where: n denotes the number of observations in the dataset y_i: Actual value of the target variable for the ith observation $\hat{y}_i$: Predicted value of the target variable for the ith observation
Mean Squared Error (MSE):	MSE computes the average squared difference between the forecasted and actual electricity consumption values. Lower MSE values indicate better accuracy.	$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2$$ Equation 3.3: Mean Squared Error Where: n denotes the number of observations in the dataset y_i: Actual value of the target variable for the ith observation $\hat{y}_i$: Predicted value of the target variable for the ith observation
Root Mean Squared Error (RMSE):	RMSE calculates the square root of the average squared difference between the forecasted and actual electricity consumption values (Hodson, 2022). A lower RMSE values indicate better accuracy.	$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2}$$ Equation 3.4: Root Mean Squared Error Where: n denotes the number of observations in the dataset y_i: Actual value of the target variable for the ith observation $\hat{y}_i$: Predicted value of the target variable for the ith observation
R^2 Score	The R^2 Score (Coefficient of Determination) quantifies the proportion of variance in the dependent variable that can be explained by the independent variables. It ranges from 0 to 1, with higher values indicating better model performance.	$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y})^2}$$ Equation 3.5: Mean Squared Error Where: n denotes the number of observations in the dataset y_i: Actual value of the target variable for the ith observation $\hat{y}_i$: Predicted value of the target variable for the ith observation $\bar{y}$: Mean of the actual values of the target variables.

Mean Absolute Percentage Error (MAPE):	MAPE measures the average percentage difference between the forecasted and actual electricity consumption values. It provides a relative measure of accuracy in terms of percentage difference two values. A lower MAPE indicates higher accuracy, as it means the average percentage difference between predicted and actual values is smaller.	$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{yi - \hat{y}i}{yi}\right\rvert$$ Equation 3.6: Mean Absolute Percentage Error Where: n denotes the number of observations in the dataset y_i: Actual value of the target variable for the ith observation $\hat{y}_i$: Predicted value of the target variable for the ith observation

3.7.6 Model hybridisation

This process provides a structured approach to integrating machine learning and statistical models, ensuring that the final hybrid model leverages the unique capabilities of each base model. In this study, a stacking ensemble method (detailed in section 2.6 of the previous chapter) was employed for hybridisation, where two meta-learner models – Linear Regression and Gradient Boosting Regressor as explained in

Table 3. 8 – were used to combine the predictions of the top performing base models.

Table 3. 8: Description of the meta-learner models used to hybridise

Meta-Learner Model	Description
Linear Regression	According to (Kpomonè et al. (2024) Linear Regression models the relationship between the dependent variable (actual electricity consumption) and one or more independent variables (predictions from the two best performing models) using a linear equation. It is computationally efficient and easy to implement.
Gradient Boosting Regression	GBR constructs multiple decision trees successively, each tree rectifies the errors of the previous one by minimizing a loss function. It captures complex, non-linear relationships in the data.

The stacking ensemble approach involves training a meta-learner on the predictions of the base models (Random Forest and ARIMA) to create a more accurate final prediction. This method allows the hybrid model to benefit from the strengths of both machine learning and statistical techniques.

The top two performing models from both the Machine Learning and Statistical categories were hybridised to leverage their combined strengths for the forecasting task. The implementation of both hybrid models follows a similar process, which is detailed below. Refer to Appendix D.9 for the model hybridisation code snippet.

Combining predictions: The predictions from both the Random Forest Regressor and the ARIMA model are combined into a new dataset. This is achieved by stacking the inverse-transformed predictions from the Random Forest and the predictions from the ARIMA model side by side. The resulting array has each row containing a prediction from the Random Forest and the corresponding prediction from the ARIMA model.

Training the meta-learner: For the Linear Regression hybrid, a Linear Regression model is used as the meta-learner. For the Gradient Boosting Regression hybrid, the Gradient Boosting Regressor (GBR) is used as the meta-learner. The meta-learners are trained on the combined predictions to learn how to best combine the predictions from the two models to minimise the error.

Mathematical expression for the hybrid model: The predictions from both models are stacked into a dataset:

$X_{meta} = [\hat{Y}_{RF} \ \ \hat{Y}_{ARIMA}]$

Where:

$\hat{Y}_{RF}$ is the predicted electricity demand from **Random Forest**

$\hat{Y}_{ARIMA}$ is the predicted electricity demand from **ARIMA**

X_{meta} is the new feature set used to train the meta-learner

The meta-learner is trained to learn an optimal weighted combination of the two models' predictions.

For **Linear Regression (LR) Hybrid Model**:

$\hat{Y}_{hybrid} = w_1\hat{Y}_{RF} + w_2\hat{Y}_{ARIMA} + \epsilon$

For **Linear Regression (LR) Hybrid Model**:

$\hat{Y}_{hybrid} = GBR(X_{meta})$

Where:

$w_1, w_2 =$ weight parameters learned by Linear Regression meta learner

$GBR(X_{meta})$ = prediction function of the Gradient Boosting Regressor

ϵ = residual error term

Final predictions: After training, the meta-learners make the final predictions on the same test period. These predictions are then evaluated against the actual values using the same evaluation metrics.

3.7.7 Electricity consumption forecast

The best performing hybrid model was then utilised to forecast the electricity consumption for the City of Windhoek for the next two years to fulfill the main objective of the study.

By combining the predictions of the Random Forest and ARIMA models and using GBR (best performing hybrid) as a meta-learner, the approach leverages the strengths of both models to provide accurate forecasts. The implementation involved generating future sequences, making predictions with individual models, and combining these predictions to produce the final forecast. Appendix D.10 shows the code snippet detailing the implementation for forecasting electricity consumption using the best performing hybrid model.

The forecasting model development process is summarised in Figure 3. 9 below.

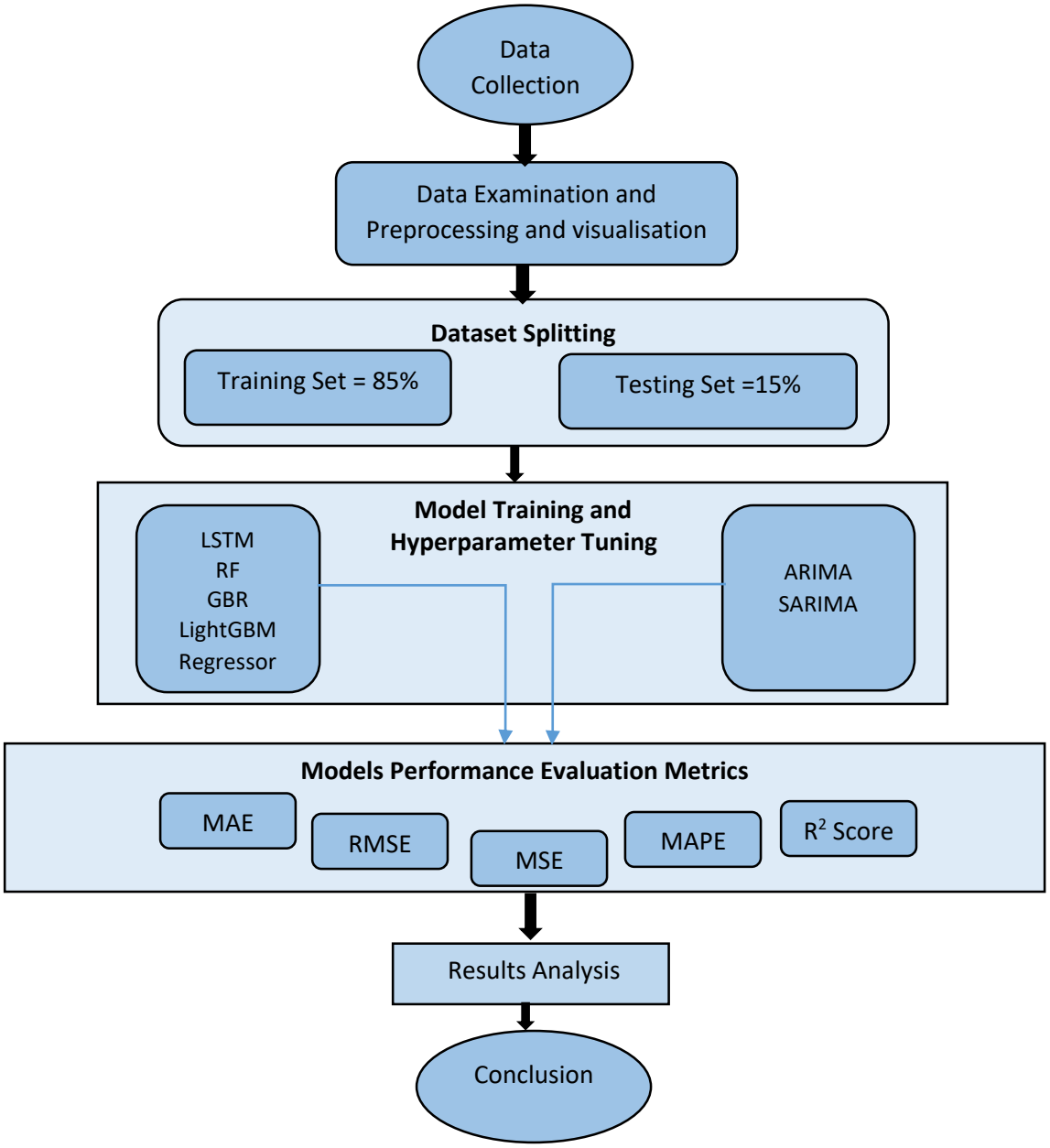


Figure 3. 9: Model development process. Source: author

3.7.8 Results analysis

After the model evaluation process, the results were thoroughly analysed to draw meaningful conclusions. This analysis involved examining the performance of each model based on the evaluation metrics such as MAE, RMSE, and MAPE. The detailed results of the model evaluations, along with their interpretations, are presented in Chapter 4, where the findings are discussed.

Additionally, insights gained from data visualisation were incorporated into the analysis to provide a broad understanding of the forecasting process. The results were carefully interpreted to identify patterns and trends within the data, providing perceptions into the correlation between various features and the forecasted electricity consumption.

The qualitative data collected from the surveys of the energy expects and policy makers was also analysed. Each participant's response was individually examined to extract pertinent information. Subsequently, all responses were collectively analysed to identify common themes as well as unique insights emerging from the data. This approach aligns with the emphasis on understanding the underlying meaning within the qualitative data, as highlighted by Gibson (2019).

3.8 CHAPTER SUMMARY

This chapter highlights the methodology employed to address the research aims and objectives as outlined in Chapter 1. The main objective guiding this study is to apply a hybrid forecasting method that combines both the traditional time series methods and machine learning algorithms for electricity demand forecasting for the City of Windhoek. Table 3. 9 below encapsulates and summarises the key components to achieve these objectives.

Table 3. 9: Summary of research objectives and evaluation criteria

Research sub-objectives	**Key tools/instruments**	**Data and key analytics**	**Expected deliverables**
To explore the trends in electricity demand for the City of Windhoek for the past ten years.	- Python Libraries (Pandas, Matplotlib, Scipy, Seaborn)	- Cow Historical Electricity Consumption Data - Data Visualization - Trend Analysis	- Identification and summary of key trends and patterns in electricity demand from 2012 to 2022.
To analyse the relationship between electricity demand and factors thought to influence electricity demand, such as GDP, population growth and renewable/alternative energy integration.		- Cow Historical Electricity Consumption Data - Economic Data (GDP, Population) - Renewable Energy Data - Correlation Analysis	- Insights into the relationship between economic indicators, population growth, renewable energy integration, and electricity demand.
To analyse the variance between the Notified Maximum Demand (NMD) and the Actual Demand (AD) for CoW, and assess the financial implications of any disparities.	- Python Libraries (Pandas, Numpy)	-- NMD Data - AD Data - Financial Data	- Analysis of discrepancies between notified and actual demand. - Financial implications of these discrepancies.

		- Variance Analysis - Financial Impact Assessment	
To evaluate the performance of various machine learning and statistical models to determine the most accurate model for forecasting electricity demand.	- Machine Learning Models (Random Forest, Gradient Boosting, LightGBM) - Statistical Models (ARIMA, SARIMA) - Grid Search CV	- Model Training and Evaluation - Performance Comparison - Hybrid Model Development	- Comparative analysis of model performances. - Identification of the most accurate models. - Development and assessment of hybrid models.
To forecast the monthly electricity demand for the City of Windhoek for the next two years.	- Best Performing Models - Hybrid Model	- Forecasting - Time Series Analysis	- Monthly electricity demand forecast for the next two years. - Analysis of demand trends over the forecasted period.
To analyse energy experts' perceptions alongside historical data to assess key factors impacting the electricity distribution industry.	- Survey Questionnaire	- Survey Data (Energy Experts' Responses)	- Qualitative and Quantitative Analysis

CHAPTER FOUR

RESULTS AND DISCUSSION OF FINDINGS

4.1 CHAPTER OVERVIEW

This chapter presents the results and discusses the findings of the study, discussing the implications and significance of the results in relation to the research objectives. The chapter outline is shown in Figure 4. 1 below.

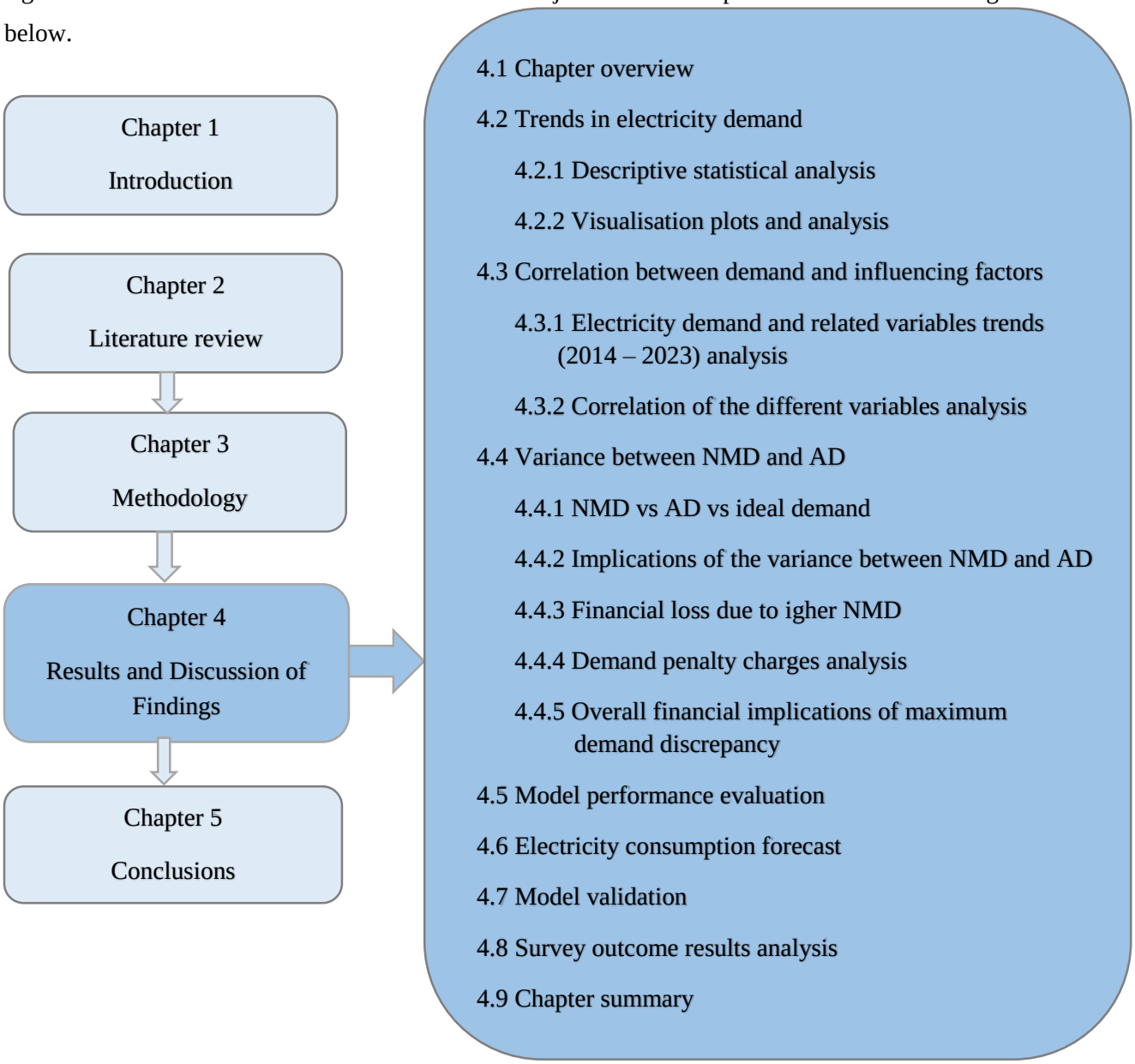


Figure 4. 1: Chapter 4 outline map. Source: author

4.2 TRENDS IN ELECTRICITY DEMAND

This section explores the electricity consumption trends over the past decade using data visualisations to highlight key patterns. These trends are critical for understanding the changes in electricity demand within the city

4.2.1 Descriptive statistical analysis

The first step was to compute the statistical summary for numerical variables in the dataset using the describe function as shown in Figure 4. 2 below. These results only depict the variables that were deemed relevant to the objectives of this study, as the columns that do not significantly contribute to the analysis of electricity demand forecasting, economic trends, population growth, and renewable energy integration were dropped in the data preprocessing process.

```
# Display descriptive statistics for the cleaned dataset
print(cow_cleant_data.describe())
```

```
       PF_@_Max.Dem  Max._Demand_in_kVA  Energy_Peak_in_kWh  \
count    123.000000          123.000000        1.230000e+02
mean       0.982342       153497.585082        2.008950e+07
std        0.008802         8454.292477        3.536508e+06
min        0.940900       131904.000000        1.315926e+07
25%        0.979150       148037.860000        1.632852e+07
50%        0.983400       153497.585100        2.134857e+07
75%        0.988650       159015.000000        2.301714e+07
max        0.998400       174585.780000        2.639814e+07

       Energy_Standard_in_kWh  Energy_Offpeak_in_kWh  Total_Energy_in_kWh  \
count            1.230000e+02           1.230000e+02         1.230000e+02
mean             2.573205e+07           3.037116e+07         7.619271e+07
std              2.332201e+06           1.791797e+06         4.398563e+06
min              1.867212e+07           2.525082e+07         6.161087e+07
25%              2.405190e+07           2.925999e+07         7.354885e+07
50%              2.548694e+07           3.033540e+07         7.596990e+07
75%              2.713020e+07           3.156784e+07         7.931213e+07
max              3.200668e+07           3.513429e+07         8.538539e+07

       Total_Energy_Amount_(N$)  Total_Amount_for_the_month_(N$)
count              1.230000e+02                     1.230000e+02
mean               8.467483e+07                     1.352686e+08
std                1.829827e+07                     2.714280e+07
min                4.333988e+07                     7.114483e+07
25%                7.470060e+07                     1.216142e+08
50%                8.227833e+07                     1.367046e+08
75%                9.350227e+07                     1.522627e+08
max                1.349437e+08                     1.979268e+08
```

Figure 4. 2: Descriptive statistical analysis. Source: author

The descriptive statistic provides insights into the characteristics and the distribution of variables of the data at hand. These statistical characteristics in general and with respect to the different variables are further explained in Table 4. 1 and Table 4. 2, respectively.

Table 4. 1: General descriptive statistical characteristics

Statistic Characteristics	**Data description and insights**
Count	These are the number of observations (data points) for each variable available for analysis which is 123 for every feature in the dataset further indicating that there are no missing values in the cleaned dataset.
Mean	This is the average value of each variable and a measure of central tendency, it provides a central value around which the data points are distributed.
Standard deviation	This is a measure of dispersion or variability indicating how spread out the data points are the mean (Popenoe et al, 2021). A higher standard deviation means that the data points are more spread out, indicating greater variability. In contrast, a lower standard deviation indicates that the data points are closer to the mean, indicating less variability.
Minimum	This is the lowest observed value for each variable, giving insight into the lower bound of the data range.
Maximum	This is the largest observed value for each variable in the dataset, giving insight into the upper bound of the data range.
Percentiles	**25th and 75th Percentiles**: According to (Popenoe, Langius-Eklof, & Jervaeus, 2021) the 25th and 75th Percentiles represents the value below which 25% and 75% of the data falls respectively. These metrics help understand the spread and skewness of the data. **Interquartile Range (IQR)**: The difference between the 75th and 25th percentiles (IQR) indicates the spread of the middle 50% of the data.

Table 4. 2: Statistical descriptive analysis

Variables	**Statistical analysis**
Power factor	The average power factor is 0.982, with a small standard deviation (0.0088), indicating that the power factor is consistently high across the observations. A high power factor suggests an efficient electrical power network.
Maximum demand:	The maximum demand varies significantly, with a mean of 153497.59 kVA and a standard deviation of 8454.29 kVA. The wide range between the minimum (131904

	kVA) and maximum (174585.78 kVA) values indicates that the city's electricity demand fluctuates considerably over time.
Peak Energy Consumption:	The average peak energy consumption is 20089500 kWh, with considerable variability (standard deviation of 3536508 kWh). The minimum and maximum values (13159260 kWh and 26398140 kWh, respectively) show substantial fluctuations in energy consumption during peak periods.
Standard and Off-peak Energy Consumption:	Similar variability is observed in standard and off-peak energy consumption, reflecting the different usage patterns during these periods.
Total Energy Consumption	The total energy consumption averages 76192710 kWh, with a standard deviation of 4398563 kWh, indicating significant variations in total monthly energy consumption.
Financial Implications	The average monthly cost of energy consumption is 84674830 N$, with total monthly bills averaging 135268600 N$. The high standard deviation for these financial figures indicates considerable variability in monthly energy costs, which could be influenced by fluctuations in consumption patterns, peak demand charges, and other factors.

4.2.2 Visualisation plots and analysis

The following plots were used for the purpose of data visualisation and to draw insights from the data analysis.

(a) Distribution of total energy consumption

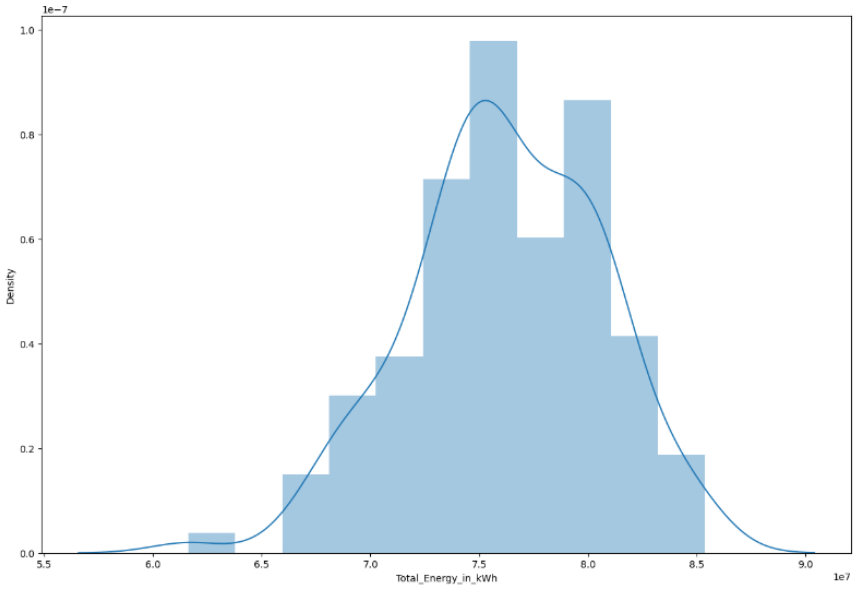


Figure 4. 3: Distribution of total energy consumption. Source: author

Figure 4. 3 above supports the descriptive statistic analysis for the total energy consumption which was the target variable for this study. It provides insights on the central tendency, variability and distribution of the variable.

The histogram and density plot show that the total energy in kWh data is approximately normally distributed with a slight right skew. The central tendency is around 7.0 to 7.5 million kWh, with most data points falling within the range of 6.0 to 8.5 million kWh. This means that while there is some variability in energy consumption, it is relatively stable, with most months exhibiting energy usage close to the central tendency. The slight skewness to the right indicates that there are occasional periods of higher energy consumption, but these are not the norm. This could be due to seasonal factors, unusual events, or increased demand in specific months. The overall shape of the distribution implies that the energy demand forecasting model can be optimised for accuracy within the typical consumption range, but it also needs to account for the potential for higher-than-average consumption in certain periods.

(b) Total energy consumption trends

Figure 4. 4 below shows the Total Electricity Consumption Trends for the City of Windhoek for the past ten years ranging from January 2014 to March 2014. The figure shows the overall energy consumption trends, representing total kWh consumed each month.

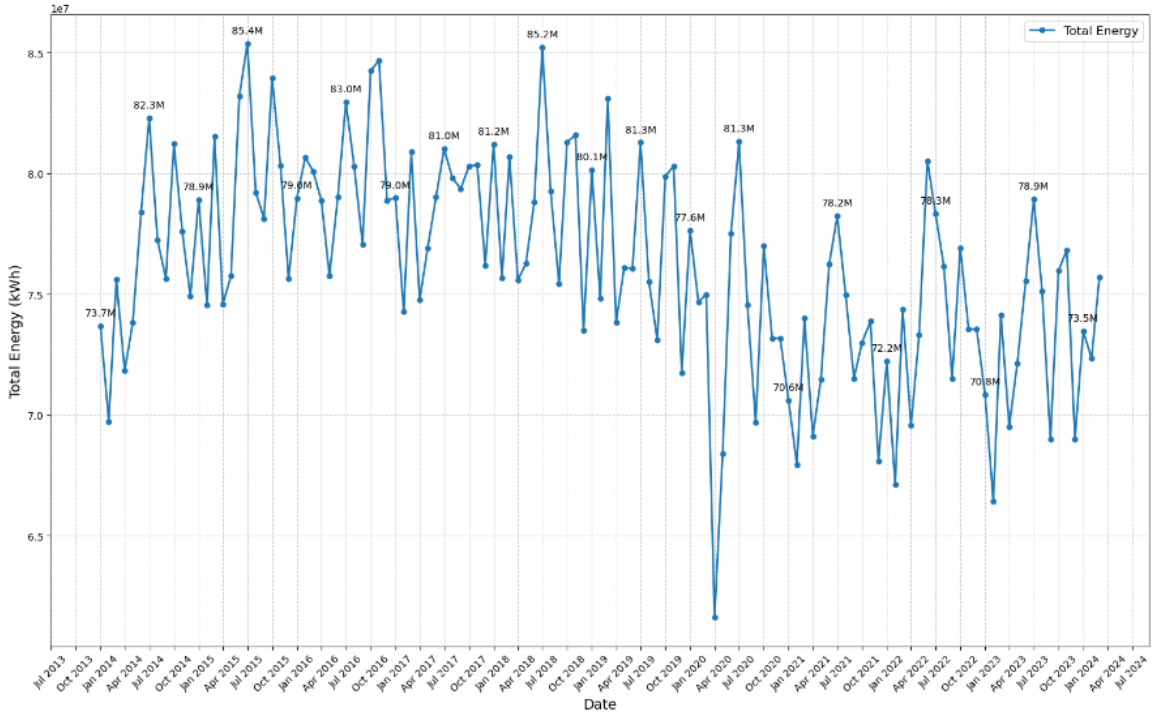


Figure 4. 4: Total energy consumption trends. Source: author

Insights from the plot of total electricity consumption trends

Overall Trend: The total electricity consumption in the City of Windhoek shows a cyclical pattern with noticeable fluctuations over the observed period. There are recurrent peaks and dips in the data indicating seasonal variations affecting the electricity consumption with the higher consumption being observed in the months June-August every year likely due to heating demands as these are the winter months in Namibia. However, a study done by Kolawole et al. (2017) on electricity demand in Sub-Saharan Africa contrasts these findings, showing that in some regions, economic activities rather than seasonal variations are the primary drivers of electricity consumption. This difference highlights the impact of local climatic and economic conditions on electricity consumption trends.

Increasing trend (2014 - 2017): From July 2013 to early 2017, there is a general upward trend in electricity consumption. This could be due to population growth, increased economic activities, or higher energy needs in the city.

Stabilization and slight decline (2017 - 2020): Between 2017 and early 2020, the electricity consumption appears to stabilise with slight fluctuations around a relatively high average. This period could indicate a phase where the energy demand was consistently high but not growing significantly.

Notable decline and variability (2020 - 2024): A significant decline in electricity consumption is observed around early 2020, followed by increased variability. This period corresponds with the global COVID-19 pandemic, which likely impacted economic activities, population behaviours, and consequently, electricity consumption.

Post-pandemic recovery (2021 - 2024): Post-2020, the data shows increased variability with no clear trend of recovery to pre-2020 levels. This could indicate ongoing economic uncertainties, changes in energy efficiency measures, or shifts in population dynamics and industrial activity.

(c) Time of use electricity consumption trends

Figure 4. 5 below depicts the Time of Use Electricity Consumption Trends for the City of Windhoek for the study period. The timeline is shown on the x-axis, whereas the y-axis shows the energy consumption in kWh. The figure shows total consumption disaggregated by time-of-use categories, namely Peak Energy (red solid line), Standard Energy (blue dashed line), and Off peak Energy (green dotted line), which provides a more detailed view of how consumption varies based on time of day and tariff structures.

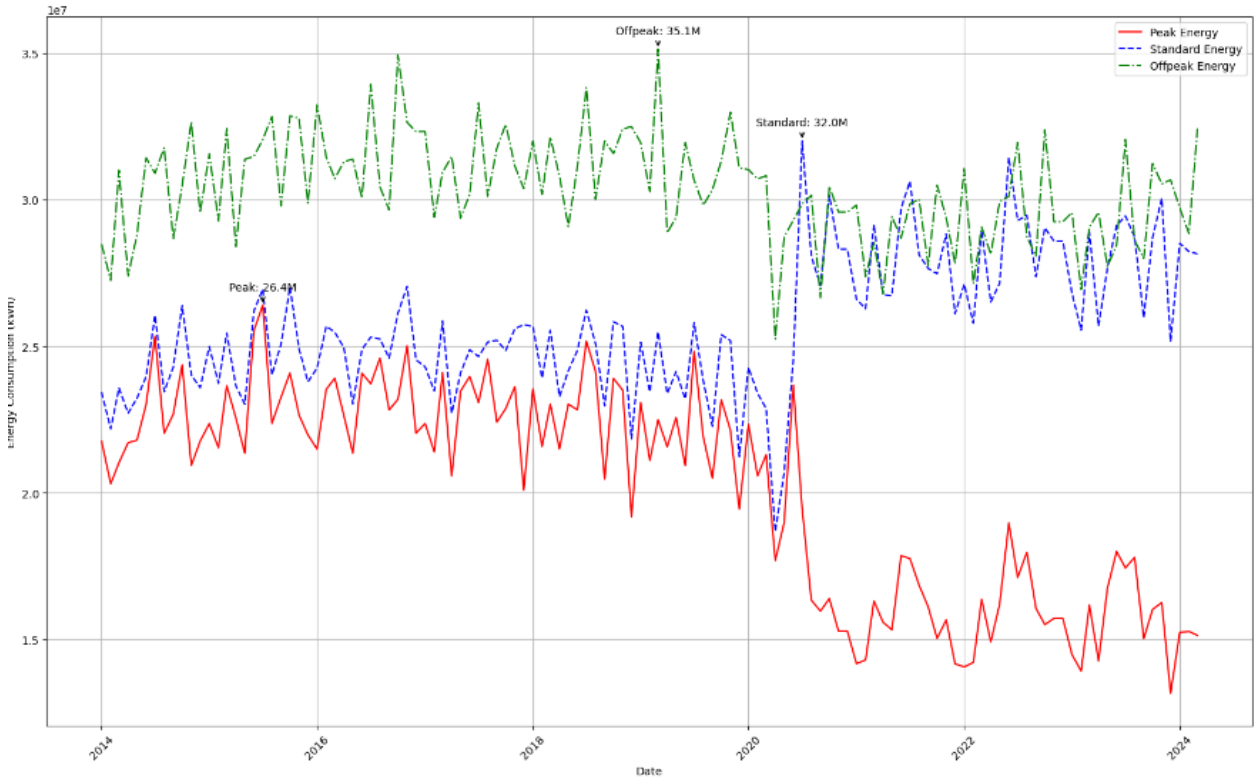


Figure 4. 5: Time of use electricity consumption trends. Source: author

The City of Windhoek's time-of-use energy structure categorizes different hours of the day based on varying levels of electricity demand, with corresponding differences in tariff rates. Peak hours, which generally occur during times of high demand (typically afternoons and early evenings), come with the highest electricity tariffs. Off-peak hours generally correspond to late evening, night, and early morning periods, when overall demand is lower, and electricity costs are significantly reduced. Consumers strategically reduce usage during costly peak periods and maximize consumption during more affordable off-peak times. This pricing structure explains why Off-peak Energy consumption consistently exceeds Peak Energy consumption in the graph.

The overall time of use trend reveals that peak energy consumption shows a noticeable decline starting around 2020, due to reduced peak-hour demand or improved energy efficiency, with regular seasonal peaks before 2020, typically aligning with summer months. Standard energy consumption remains relatively stable over time, with regular fluctuations and a significant spike around early 2020. Off peak energy consumption consistently surpasses both peak and standard consumption, indicating significant off-peak usage, and exhibits a stable overall trend with regular seasonal variations similar to peak energy.

(d) Box plot of total energy consumed each month

The box plot of total energy consumed each month provides a clear visual representation of the distribution of energy consumption across different months to help understand the pattern and seasonality in the data.

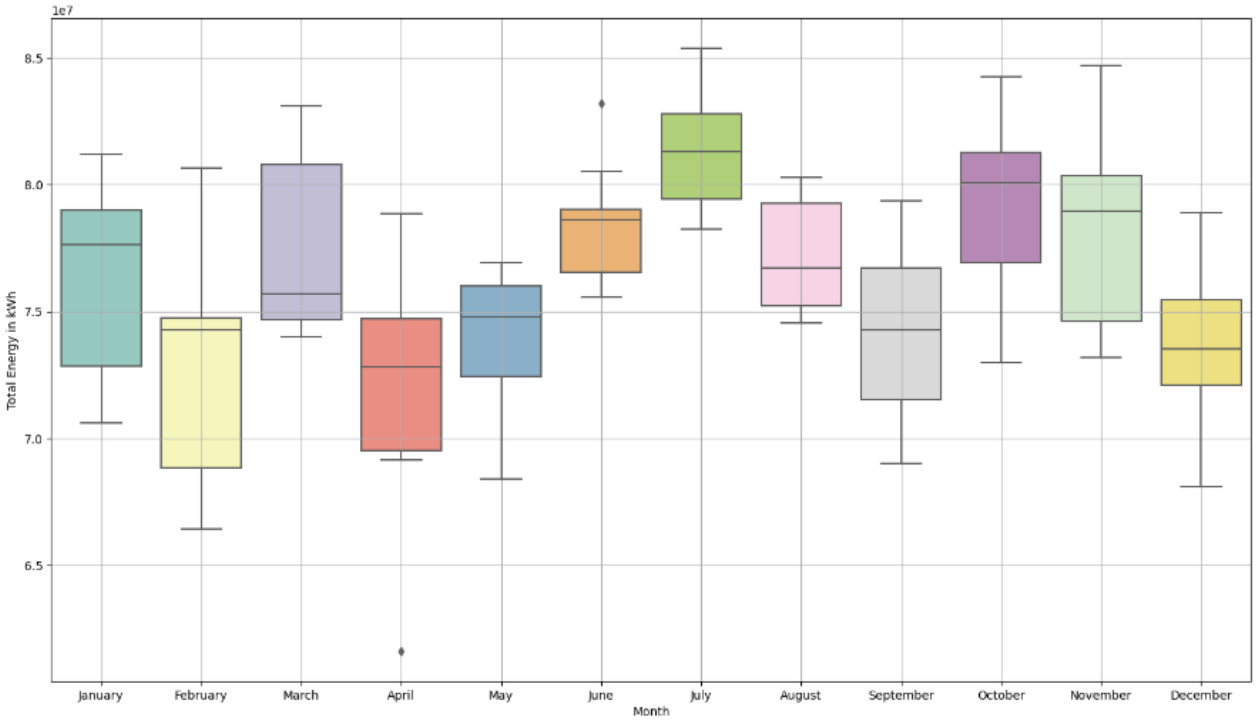


Figure 4. 6: Box plot of total energy consumed each month. Source: author

The box plot indicates clear seasonal patterns in energy consumption with an observed higher energy consumption during the winter months (June - August) which suggest an increase in heating demands. The summer months (December - February), conversely tend to exhibits a lower energy consumption, possibly due to reduced heating and cooling demands. There is, however more variability in energy consumption among the other transitional months (March - May, September - November), which could be influenced by changing weather conditions and varying energy demands. Overall, the data reveals a predictable seasonal trend with occasional anomalies, thus emphasising the importance of considering seasonality in energy demand forecasting.

(e) Seasonal decomposition of total energy consumption time series

The seasonal decomposition of the time series of the City of Windhoek total energy as shown in Figure 4. 7 results in four key components: the observed data, the trend, the seasonal component, and the residuals

which are explained below. The timeline is shown on the x-axis, whereas the y-axis shows the energy consumption in kWh.

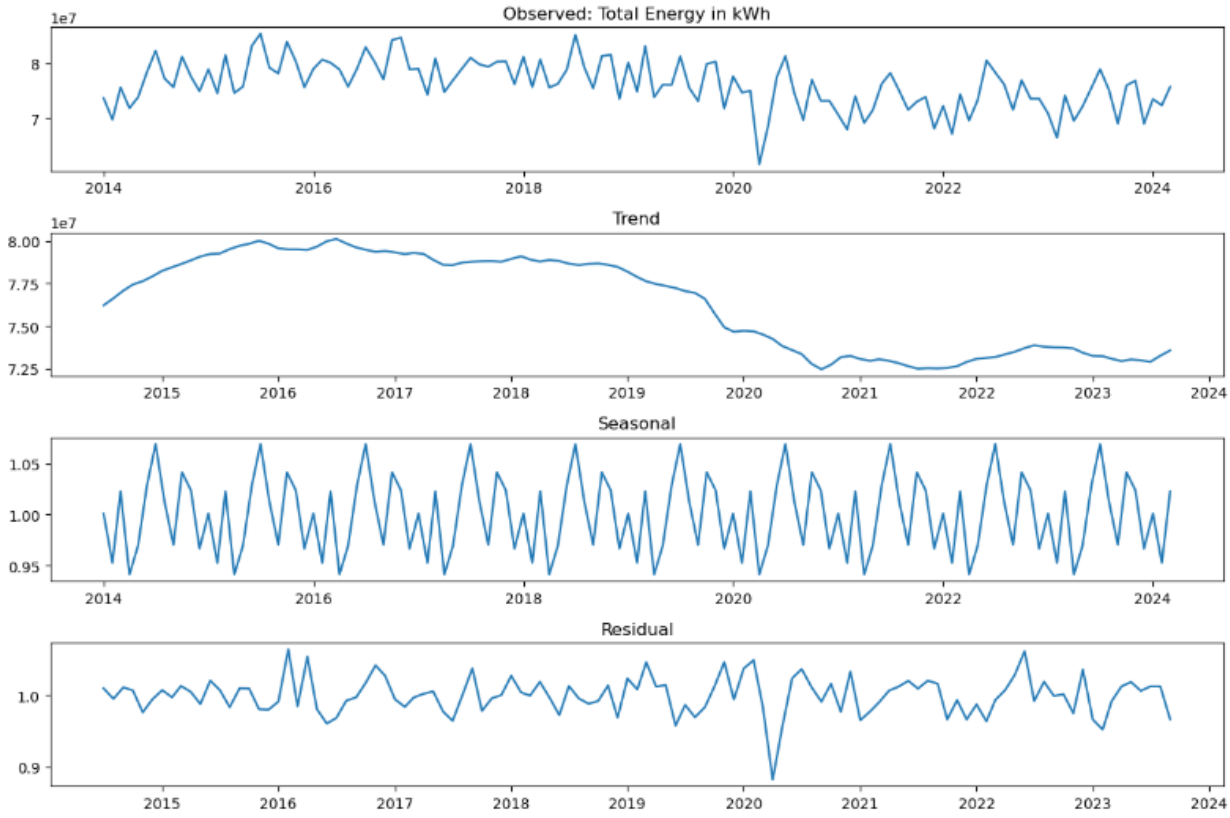


Figure 4. 7: Seasonal decomposition of total energy consumption time series. Source: author

Observed data: This is the original time series data of Total_Energy_in_kWh as shown in Figure 4. 4 with noticeable drop around the beginning of 2020 due to the COVID-19 pandemic.

Trend component: The trend component of the time series represents the long-term movement in energy consumption, excluding seasonal and irregular effects (Express Analytics, 2021). It highlights the underlying direction or pattern in the data. From 2014 to around 2019, there is a slight upward trend in energy consumption. However, starting from around 2019, the trend shows a downward movement until around the end of 2022, indicating a decrease in total energy consumption during this period. A slight upward movement is visible again towards 2024, suggesting a potential recovery or increase in energy consumption.

Seasonal component: The seasonal component of the time series captures the repeating short-term cycles, which occur at regular intervals. These cycles are typically influenced by seasonal factors such as weather

or holidays. The seasonal values are computed by averaging the data points at each corresponding season across different years, then subtracting the overall mean of the series. This component shows regular peaks and troughs that align with yearly seasonal variations in energy consumption, such as higher usage during winter months due to heating demands.

Residuals (Residual component): The residual component represents the irregular or random fluctuations in the time series after removing the trend and seasonal effects. Mathematically, the residuals are calculated using the formula:

$$\text{Residual} = \frac{\text{Observed Data}}{\text{Trend Component x Seasonal Component}} \qquad \text{Equation 4.1}$$

4.3 CORRELATION BETWEEN DEMAND AND INFLUENCING FACTORS

This section examines the correlation between electricity demand and factors such as GDP, population growth, and renewable energy integration. The analysis provides insights into how these factors impact and relate to the electricity demand.

4.3.1 Electricity demand and related variables trends (2014-2023) analysis

Figure 4.8 below illustrates the normalised trends of key factors such as electricity demand, GDP, population, and renewable energy adoption for the City of Windhoek over the study period. This combined graph allows for a comprehensive comparison of how economic activities, population growth, and renewable energy integration have influenced electricity demand in the city over the past decade.

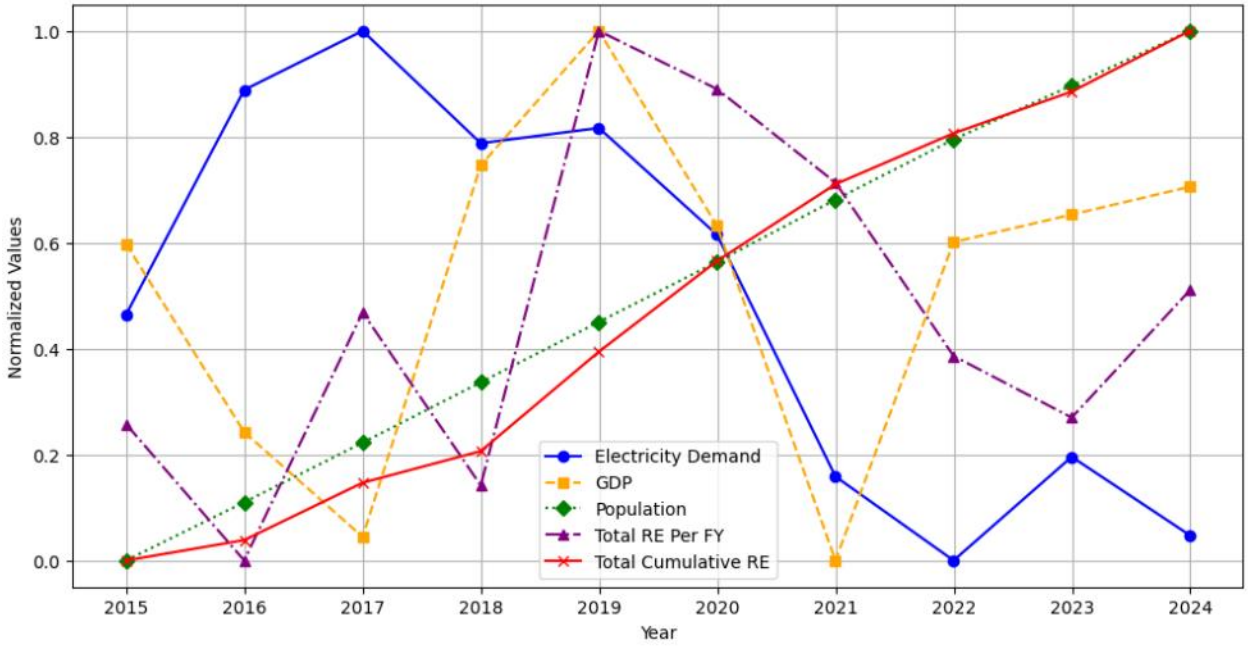


Figure 4. 8: Trends of electricity demand and related variables. Source: author

From the above graph, it is evident that there is a correlation between GDP and electricity demand trends, with both peaking around 2016-2018 and experiencing significant drops around 2020. This sharp decline in both GDP and electricity demand around 2020 may be attributed to the COVID-19 pandemic, which led to a slow-down in economic activities. This suggests that economic activities significantly influence electricity demand. It is also important to note that despite the correlation and fluctuations in GDP and electricity demand, counter-intuitively, electricity demand appears to be on a downward trajectory despite a continued growth in the population.

The steady increase in cumulative renewable energy suggests a continuous and growing adoption of renewable energy sources in Windhoek over the past decade. The observed decline in electricity demand from 2016 may be attributed to the continued increase in renewable energy adoption. This indicates that renewable energy sources are becoming a more significant contributor to meeting the city's energy needs, potentially reducing the demand from traditional electricity sources. It is important to note that both electricity demand and total renewable energy generation per financial year dropped significantly in 2020, indicating the impact of the COVID-19 pandemic. The pandemic disrupted economic activities and energy consumption patterns, leading to reduced energy demand and a temporary slowdown in renewable energy projects.

4.3.2 Correlation of the different variables analysis

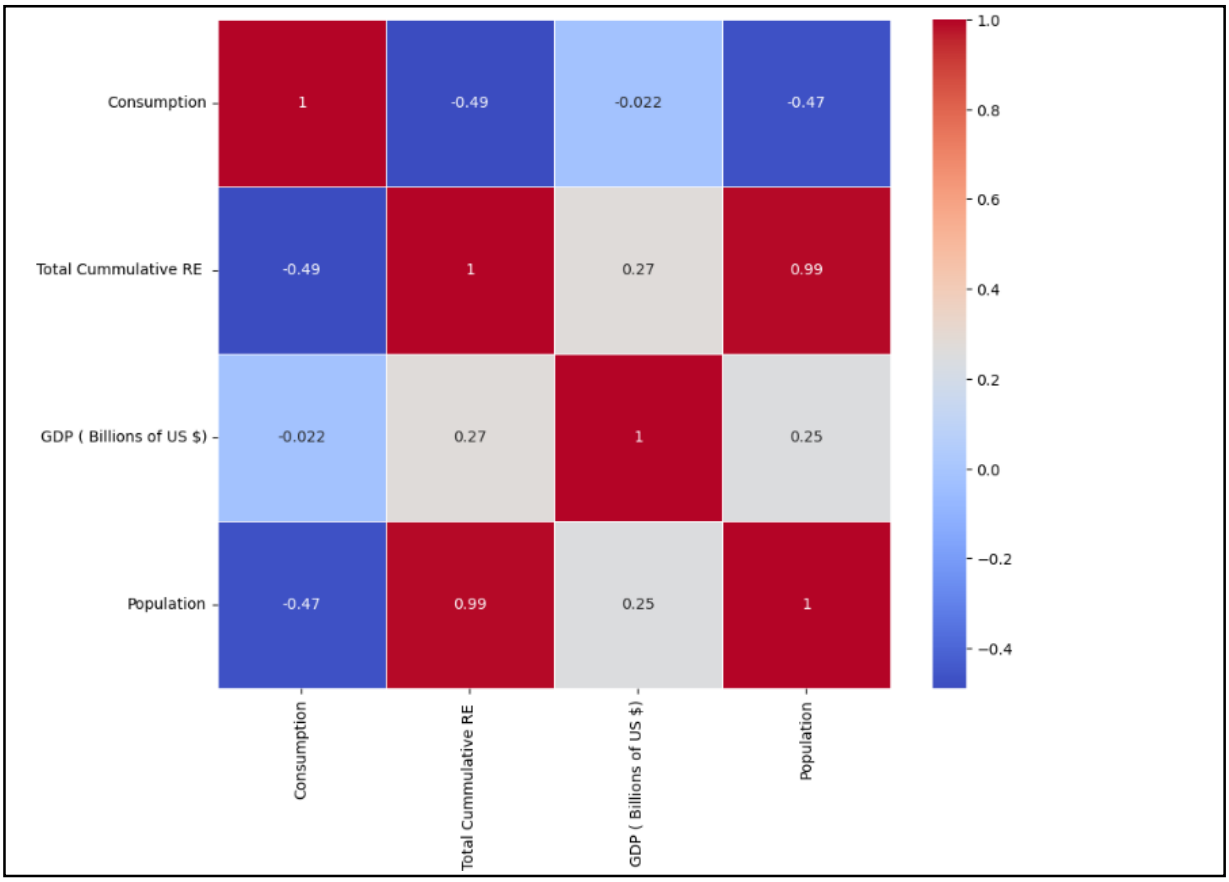


Figure 4. 9: Correlation matrix of the different variables. Source: author

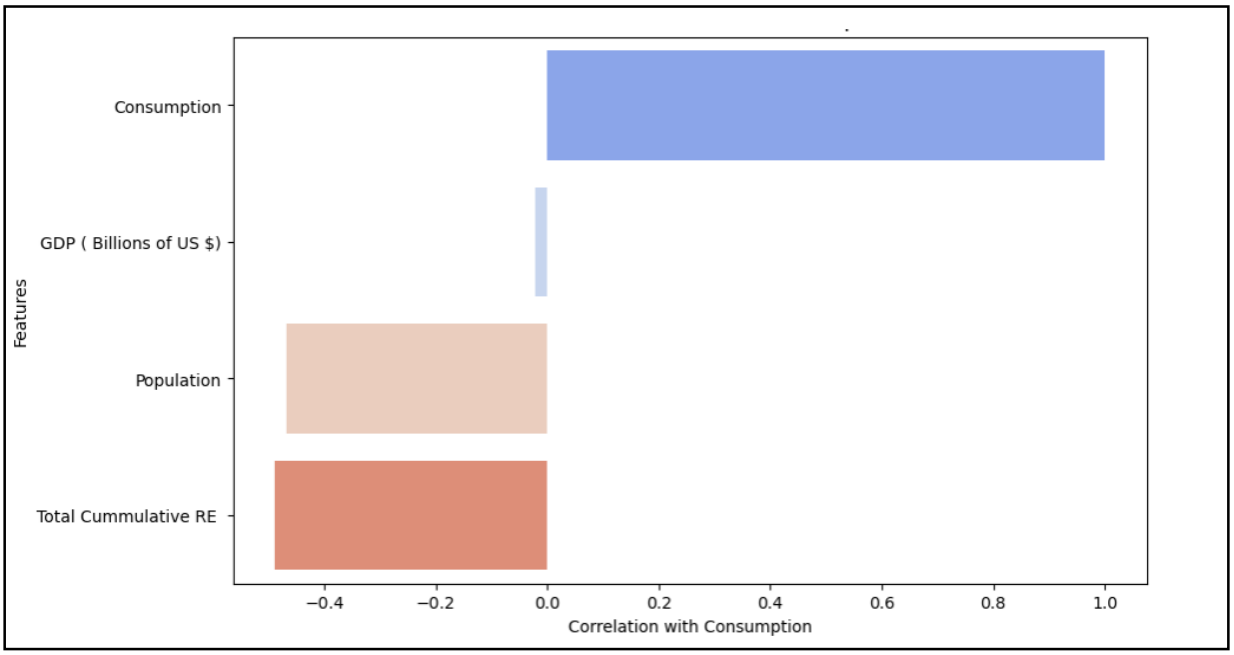


Figure 4. 10: PCA feature correlations bar plot. Source: author

The correlation matrix in

Figure 4. 9 visualises the Pearson correlation coefficients between different variables in the dataset with the correlation values ranging from -1 (perfect negative correlation) to 1(perfect positive correlation) and 0 indicates no correlation (Berman, 2016).

The strong negative correlation between Consumption and both Total Cumulative Renewable Energy and population suggests that in spite of the continued population increase which may be due to urban migration, the adoption of alternative sources of energy has been increasing at a sustained pace, resulting in a net a decline in electricity demand. Improved energy efficiency and shift towards more sustainable energy consumption practices may additionally account for the observed reduction in electricity demand from 2016.

The very strong positive correlation between Total Cumulative Renewable Energy and Population indicates that renewable energy capacity expansion is closely tied to population growth, highlighting the importance of scaling renewable energy to meet increasing energy demands.

The weak correlations between GDP and other variables suggest that GDP growth does not strongly influence nor is it influenced by the other factors.

Figure 4. 10 further illustrates the correlation of various features with electricity consumption as a results of the PCA. Total Cumulative RE indicates a negative correlation coefficient of about -0.4 in relation to the electricity consumption supporting the results of the correlation matrix.

4.4 VARIANCE BETWEEN NOTIFIED MAXIMUM DEMAND (NMD) AND THE ACTUAL DEMAND (AD)

This section presents the variance between the CoW's NMD and the actual electricity demand, and discusses the financial implications.

4.4.1 Notified Maximum Demand vs Actual Demand vs Ideal Notified Maximum Demand

The graph in Figure 4. 11 below compares the City of Windhoek's Notified Maximum Demand (NMD), Actual Demand (AD), and Ideal NMD over time from Jan 2014 to March 2024. The ideal maximum demand values, which were considered to be optimal, were determined by taking the highest maximum demand recorded within each annual cycle for comparison purposes with the annual NMD.

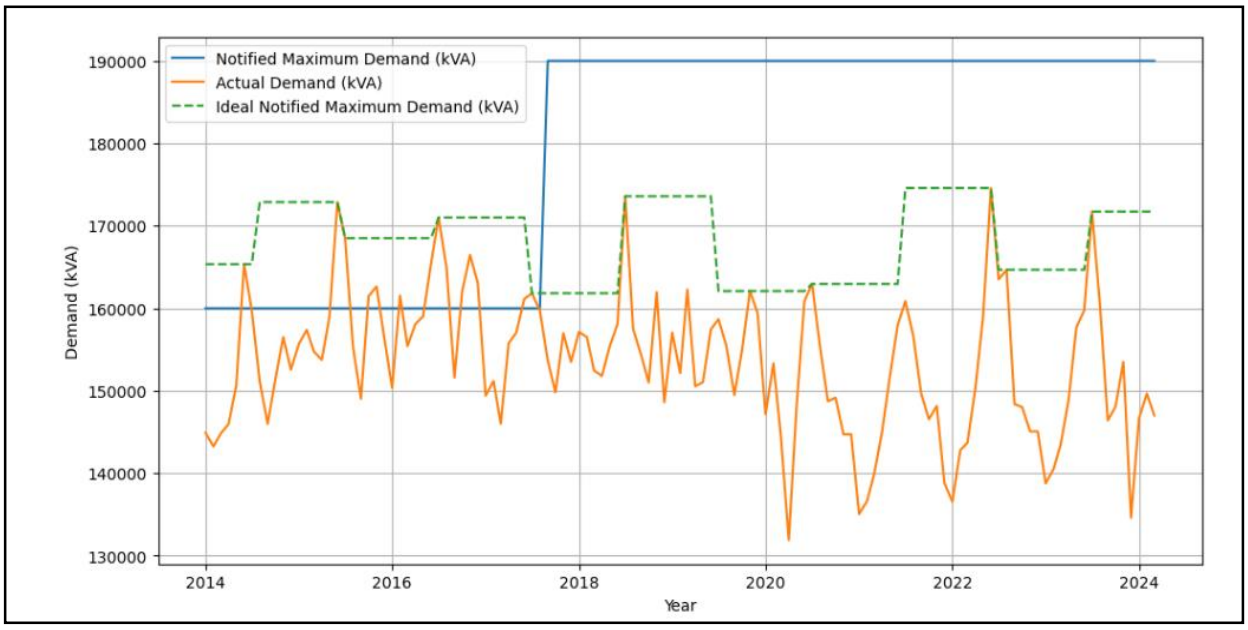


Figure 4. 11: NMD vs AD vs Ideal NMD (Source: author)

The key observation from this graph is that the NMD remained constant at 160 000 kVA from 2014 through 2016. From September 2017, the NMD changed to 190000 kVA. The actual maximum demand shows considerable fluctuations throughout the period. From 2014 to 2017, actual demand frequently exceeds the notified maximum demand, indicating periods where the City of Windhoek operated above its contracted demand. Post-2018, actual demand continues to fluctuate but way below the NMD, which had been increased to 190000 kVA. The ideal maximum demand often exceeded the notified maximum demand, particularly before 2018, highlighting discrepancies between the notified demand and actual demand. These discrepancies suggest that the initially set NMD was inaccurately determined, with the ideal maximum demand representing what the notified level should have been. From the graph, it can be seen that there is a need to improve on the NMD estimations to align with the actual demand. By adopting a more dynamic and data driven approach to setting the NMD and implementing effective demand management strategies, the city can improve the overall resource utilisation efficiency and better manage associated costs.

Figure 4. 12 further illustrates the variance between the City of Windhoek's NMD and AD to accentuate the discrepancies. The variance is categorised into positive and negative values, represented by green and red areas respectively, with black points indicating the magnitude of variance in kVA.

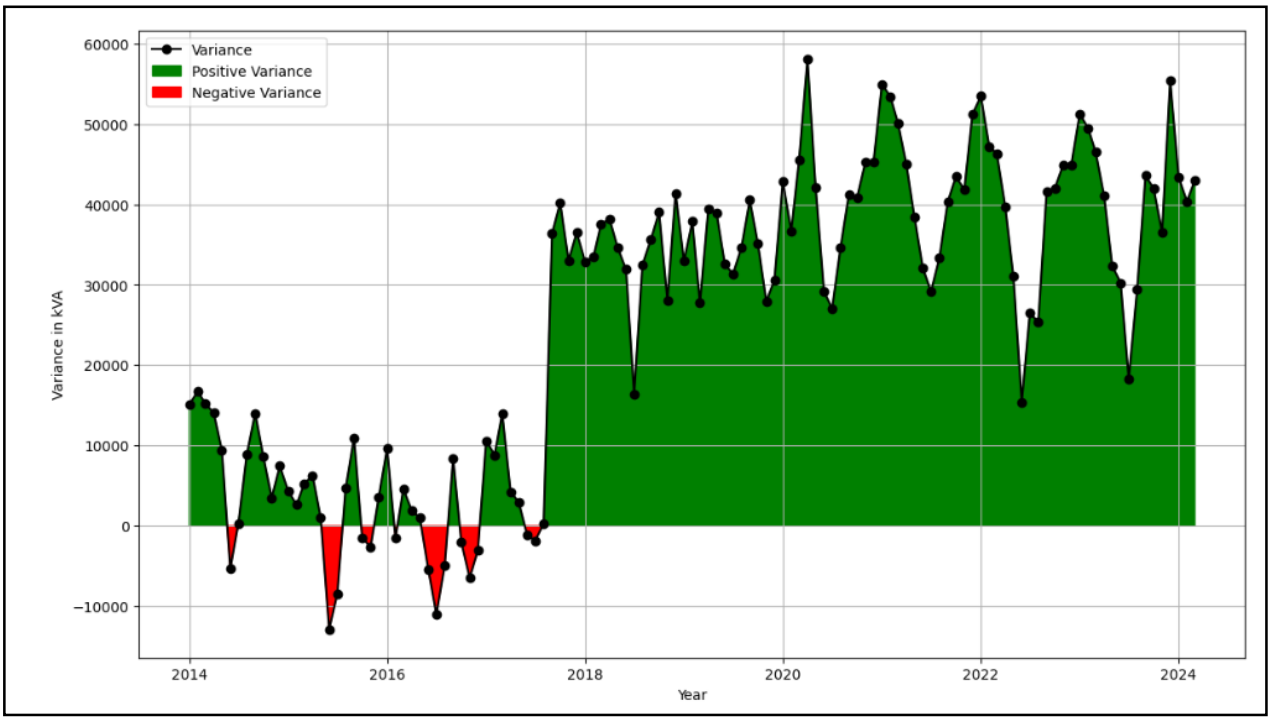


Figure 4. 12: Variance between NMD and AD (Source: author)

A positive variance specifies that the actual demand is below the notified maximum demand, specifying that the City of Windhoek paid for more capacity than it used. Conversely, a negative indicates that the actual demand exceeds the notified maximum demand, potentially leading to penalties for exceeding the contracted capacity. For about 80% of the period under study, the data shows a positive variance that increased substantially from 2018, which corresponds to the CoW's increase in the notified maximum demand to 190 000 kVA. Before 2018, there are several instances of negative variance, reflecting periods when the City of Windhoek's actual demand exceeded the notified maximum demand. The magnitude of variance shown above emphasises the need to accurately set the notified maximum demand in line with actual consumption patterns.

4.4.2 Implications of the variance between NMD and the AD

The variance between the NMD and AD in the City of Windhoek has several implications that can be broadly categorised into financial, operational, and stakeholder impacts:

(a) Financial implications

Positive variance: When the actual demand is lower than the notified maximum demand, the City of Windhoek pays for unused capacity, leading to financial inefficiencies. The unutilised kVA in one billing cycle does not roll over to the next billing cycle. Any underutilisation is billed according to the network access charge rate set for that period. Thus, the city incurs costs for the excess capacity that remains

unutilised, which could have been better allocated or managed. The financial impact of this underutilisation is significant as it contributes to unnecessary expenditure and the increased financial burden on the City of Windhoek's budget.

Negative variance: Conversely, when the actual demand exceeds the notified maximum demand, the City of Windhoek faces potential penalty charges, which are additional costs that arise from exceeding the contracted capacity. These penalties increase the overall financial outlay and indicate inefficiencies in demand management.

(b) Operational/Logistical implications:

Operational efficiency: Positive variance indicates that the city has overestimated its demand, leading to operational inefficiencies where resources are allocated based on inaccurate forecasts. This misalignment can strain the electricity supply chain, particularly in the long term, as consistent overestimation may result in redundant investments in infrastructure and capacity.

System reliability: Negative variance may challenge the reliability of the electricity grid, as exceeding the contracted demand can strain NamPower's ability to supply electricity efficiently. It can also disrupt the grid's stability, potentially leading to outages or the need for emergency measures.

(c) Stakeholder implications:

Impact on NamPower: NamPower, as the electricity supplier, guarantees the availability of the notified maximum demand. A positive variance suggests that NamPower is maintaining capacity that is not utilised, which could have been allocated elsewhere. On the other hand, a negative variance may force NamPower to activate additional capacity or source electricity at higher costs, which can affect its operations and financial planning.

Impact on CoW's clients: Positive variance, leading to higher electricity costs, might indirectly affect the tariffs charged to consumers. Negative variance could lead to service disruptions or increased tariffs to cover the penalty charges.

4.4.3 Financial losses due to higher notified maximum demand

To translate the variance between the notified maximum demand and the actual demand into monetary terms, the financial impact was quantified by focusing on the positive differences, which reflect the underutilization of the contracted electricity capacity. This monetary value was computed by multiplying the positive difference between the notified maximum demand and the actual demand (measured in kVA) by the specific network access rate applicable for each year under consideration, which for 2024 is set at

100.41 N$/kVA. This approach ensures that the financial losses are calculated accurately based on the relevant tariff rates for each specific year during the study period from 2014 to 2024.

The resulting graph, titled "Financial Loss due to Higher Notified Maximum Demand," in Figure 4. 13 visually illustrates these financial losses over time. The area under the curve represents the cumulative financial impact, highlighting periods where the underutilisation of capacity led to significant financial implications for the City of Windhoek. This financial influence is governed by the terms and conditions agreed upon between the City of Windhoek and NamPower. Under these terms, the City of Windhoek commits to a notified maximum demand, which represents the maximum electricity capacity (in kVA) that the CoW expects to require during a billing period. NamPower, in turn, guarantees to have this capacity available. However, if CoW's actual demand is lower than the notified maximum demand, the CoW still has to pay for the unused capacity based on the agreed tariff for that period.

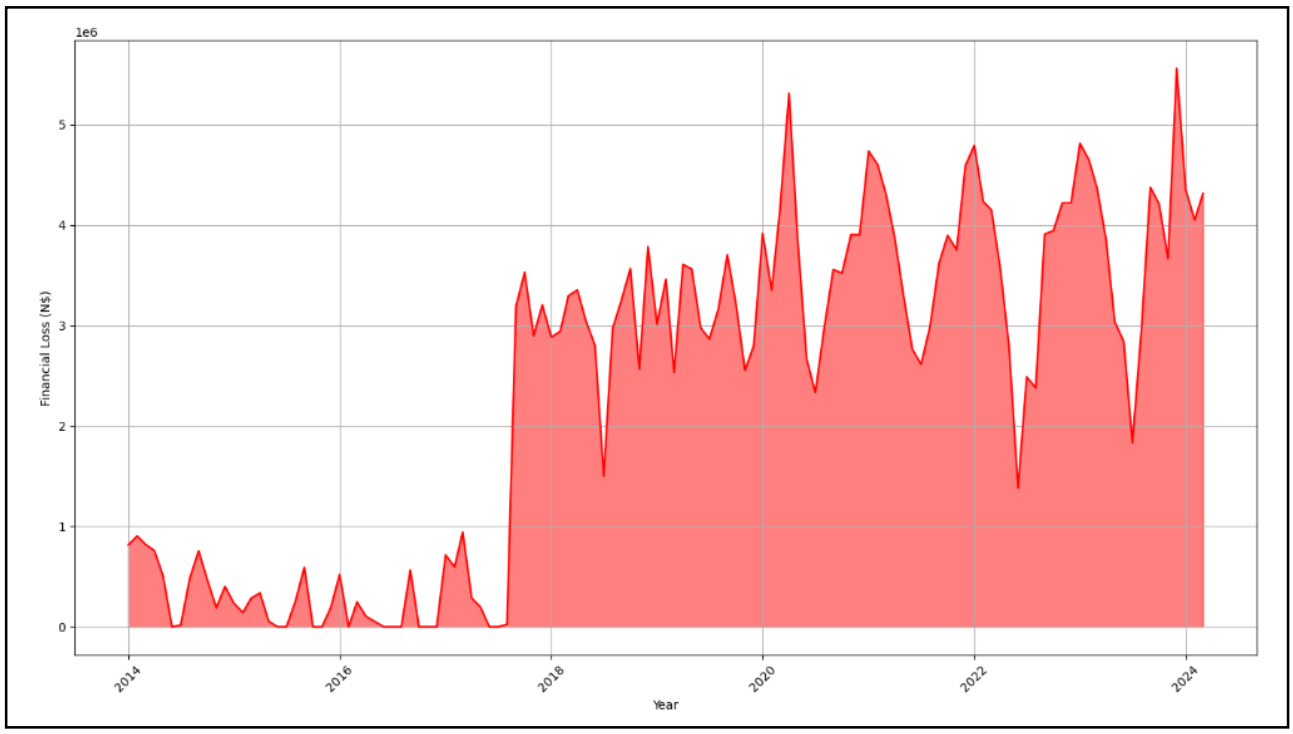


Figure 4. 13: Financial losses when the NMD is higher than the AD. Source: author

Pre-2018: The financial loss remains relatively low, with minor fluctuations. Losses are less than N$ 1 million for most of this period.

Post-2018: There is a significant increase in financial loss starting in 2018, corresponding to the increase in notified maximum demand to 190,000 kVA. Losses often exceed 3 million (N$) and reach up to around 5 million (N$) at their peak.

To contextualise the financial amounts discussed, the exchange rates as of 28 August 2024 (current date at the time of conducting this study), are approximately **1 N$ = 0.0562 US$**, **1 N$ = 0.0501 €** (Euro) and **1 N$ = 0.0427 £** (Pound Sterling).

From Jan 2014 to March 2024, the CoW recorded a total variance of 3238004.03 kVA, which translates into a total loss of N$ 288,242,177 due to maintaining a higher notified maximum demand than necessary. This implies an average monthly loss of N$ 2,402,018. These losses highlight the importance of accurately aligning the notified maximum demand with actual usage patterns to optimise electricity costs and improve resource allocation. By implementing more dynamic and precise demand forecasting and management strategies, the city can mitigate these financial losses and enhance overall efficiency in electricity consumption. Refer to Appendix D.11 for the code snippet used to calculate these financial losses.

4.4.4 Demand penalty charges analysis

Figure 4. 14 illustrates the demand penalty charges incurred by the City of Windhoek from 2014 to 2024. Penalty charges are incurred when the actual electricity demand exceeds the notified maximum demand for three consecutive months, leading to additional costs. These costs are calculated based on the total kVA variance between the notified maximum demand and the actual demand, multiplied by a penalty rate that is higher than the standard network access rate.

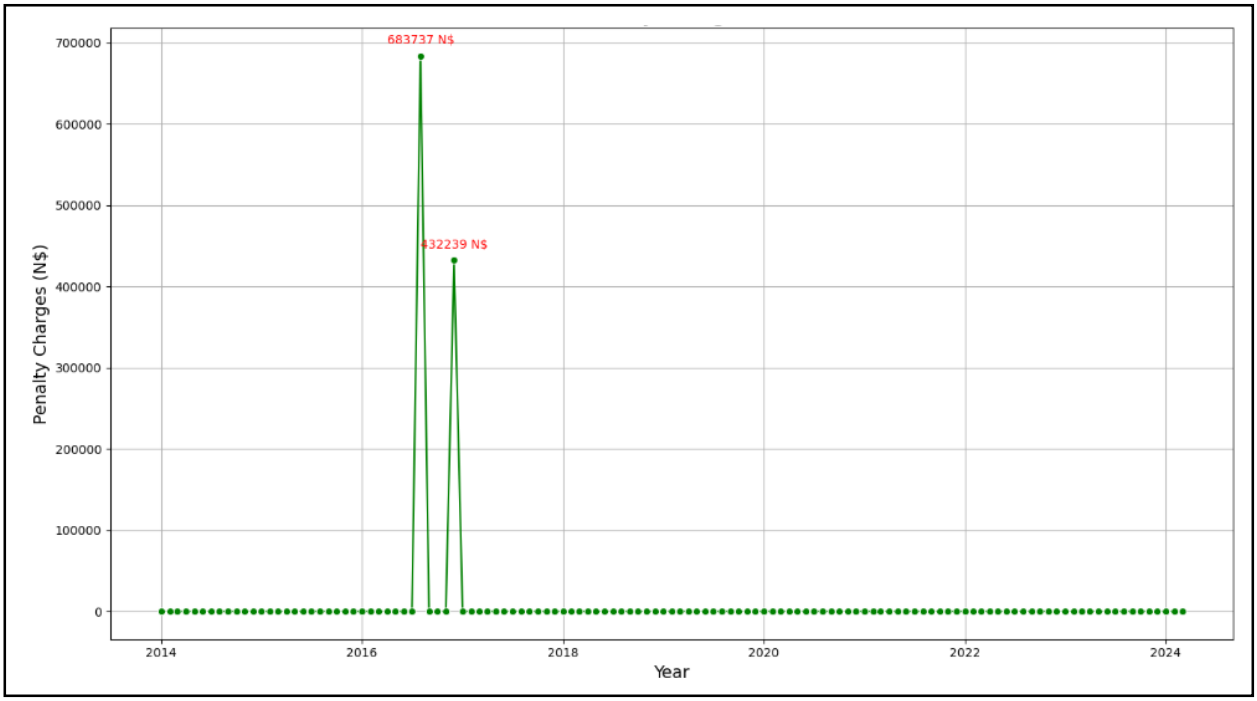


Figure 4. 14: Demand penalty charges over time. Source: author

Two significant spikes in penalty charges are observed in 2016, with values of approximately N$ 683,737 and N$ 432,239. According to the preceding evidence, the zero-penalty periods are because the CoW upped its notified maximum demand. The notified maximum demand post 2016 turned out to be way higher than the actual demand, which may be effective (from the perspective of meeting client needs), but definitely not efficient nor effective from a cost management perspective.

The observed trends suggest that while the City of Windhoek has effectively reduced penalties by increasing the notified maximum demand post-2016, it now faces the challenge of underutilisation, as the city is paying for more capacity than it needs.

4.4.5 Overall financial implications of maximum demand discrepancies

The analysis of the maximum demand discrepancies financial implications focus on three key financial metrics calculations:

1. **Total max demand charge:** This represents the total cost incurred by the City for the maximum demand actually used over the study period.
2. **Total notified demand charge:** This reflects the total cost paid for the contracted notified maximum demand, regardless of whether the actual demand reached this level.
3. **Total penalty charge:** These are penalties incurred when the actual demand exceeded the notified maximum demand consecutively for three months.

These values were also averaged on a monthly basis to provide a clearer understanding of the implications as presented in Appendix D.12.

From the computations, the total notified demand charge (N$ 1,792,427,216.42) is significantly higher than the total maximum demand charge (N$ 1,614,439,195.64). On average, the difference is about N$ 1,447,058.00 monthly. This discrepancy indicates that the City of Windhoek is consistently paying for more capacity than it actually uses, leading to higher overall costs. While the total penalty charge (N$ 1,115,976.15) and average monthly penalty charge (N$ 9,072.98) are relatively small compared to the total costs, they highlight periods where demand management was ineffective, which induced additional financial costs.

It is evident that both positive and negative variances between NMD and AD are costly for the CoW. However, based on the findings, positive variance - where AD is consistently lower than NMD, has resulted in more severe financial implications over time. This situation leads to significant financial losses due to the City paying for unused capacity. Although generally negative variance can imply potential supply risks

and penalties, the study highlights the need for the CoW to manage positive variance carefully. Over-provisioning, while avoiding the risks of under-provisioning, has proven to be costly and inefficient. Therefore, the city needs to refine its demand forecasting and adjust its NMD more dynamically to ensure that it is not excessively higher than actual demand, thereby optimising resource utilisation and minimising unnecessary financial losses.

4.5 MODEL PERFORMANCE EVALUATION

The following section outlines the analysis and comparison of results obtained from the different models explored to select the most suitable ones for the task.

4.5.1 Hyperparameter tuning analysis

(a) Machine learning models

Hyperparameter tuning is a crucial step in the machine learning model development process. It involves selecting the best set of hyperparameters that optimise the performance of a model (Bartz et al., n.d.). These parameters are set prior to the learning process and are not learned from the data. This subsection analysis focuses on comparing the performance of the models explored in this study before and after hyperparameter tuning. The key metrics used for this analysis evaluation were RMSE, R^2, and MAPE. Figure 4. 15 and Figure 4. 16 below illustrate the model performances before and after hyperparameter tuning, respectively.

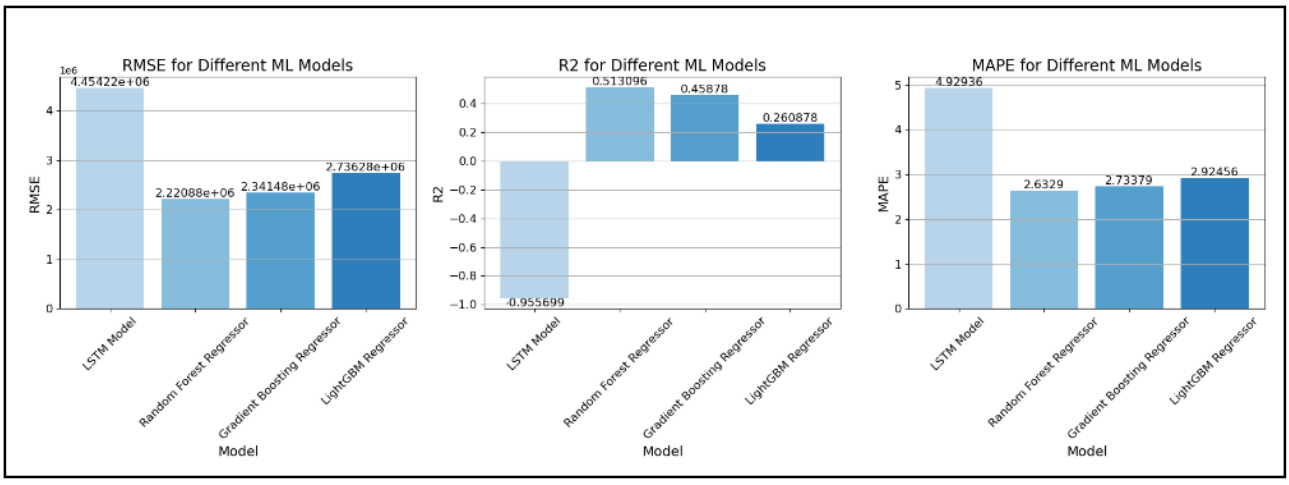


Figure 4. 15: ML models performance before hyperparameter tuning. Source: author

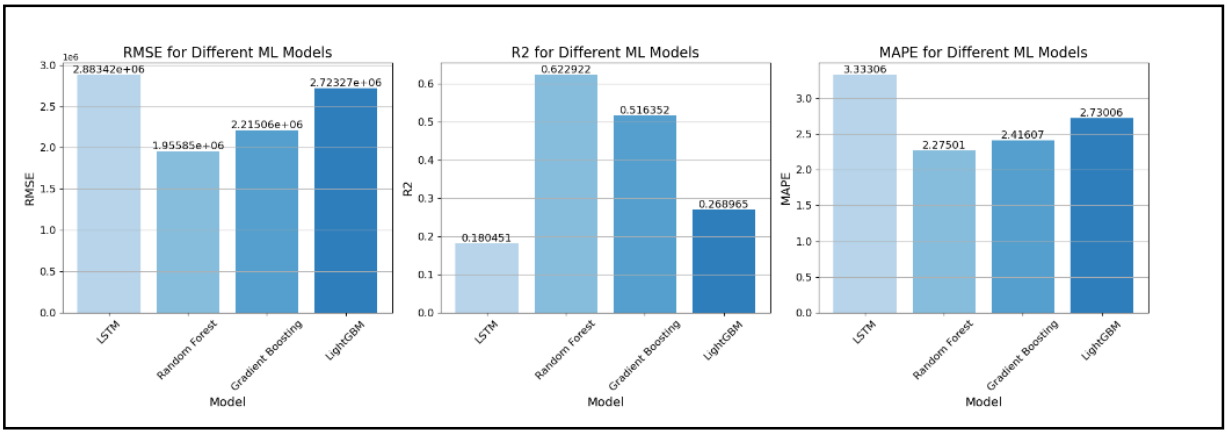


Figure 4. 16: ML models performance after hyperparameter tuning. Source: author

From the two graphs analysis it's the following key findings are evident.

RMSE for Different ML Models: The RMSE values decreased for all models after tuning, indicating better predictive accuracy.

R^2 for Different ML Models: The R^2 values increased after tuning, thus suggesting that the models were better at explaining the variance in the data.

MAPE for Different ML Models: The MAPE values decreased after tuning, indicating more accurate predictions across all models.

Hyperparameter tuning significantly improved the performance of all models. The Random Forest Regressor and Gradient Boosting Regressor showed the most notable improvements, making them the most suitable models for this forecasting task. The LSTM model, despite improvements, still underperformed compared to the other models. LightGBM also improved but remained less effective than Random Forest and Gradient Boosting models. This analysis demonstrates the importance of hyperparameter tuning in enhancing model performance and achieving more accurate predictions.

(b) Traditional statistical models

The statistical models, ARIMA and SARIMA were chosen after considering the ACF and PACF plots and ensuring stationarity through differencing and using the auto_arima respectively. The models performed better on the given dataset without further parameter adjustments, achieving the evaluation metrics results depicted in Figure 4. 17:

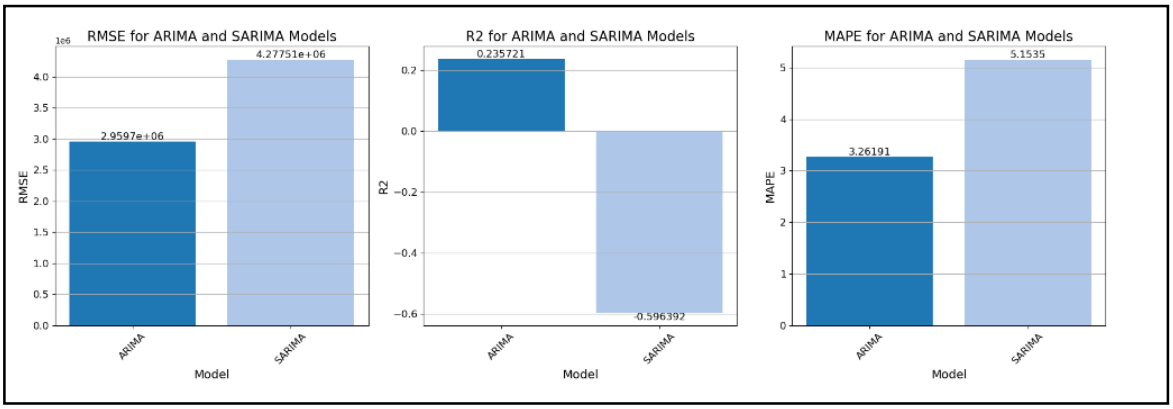


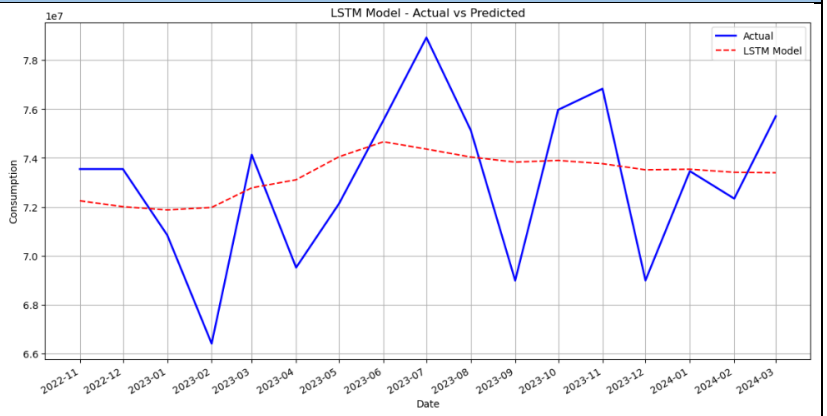
Figure 4. 17: Statistical models performance after hyperparameter tuning. Source: author

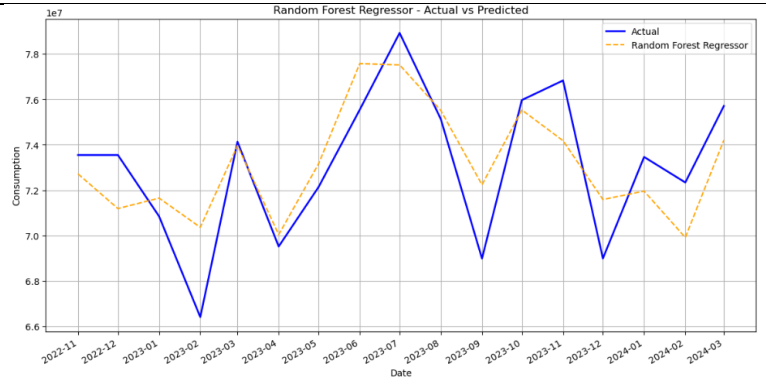
The model evaluation metrics analysed in the following sections focused on the results after hyperparameter tuning only.

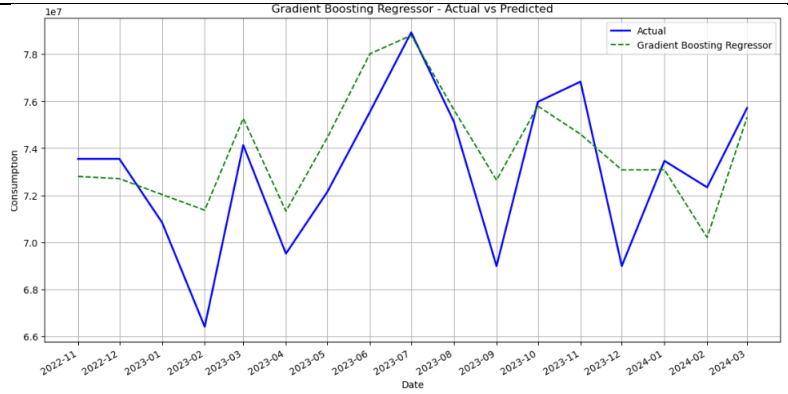
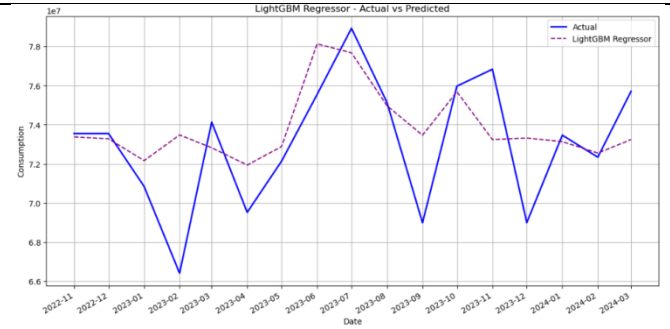
4.5.2 ML models evaluation analysis

The output results of the ML models evaluation are presented in Table 4. 3 below with the line plot of the Actual vs predicted values for the test set.

Table 4. 3: ML models evaluation analysis

Evaluation Metrics Results	Results Analysis	Actual vs Predicted graph
LSTM Model Evaluation: MAE: 2399104.61 MSE: 8314131730197.92 RMSE: 2883423.61 R^2(%): 18.045103 MAPE: 3.333059	MAE is approximately 2.4 million units of consumption, indicating that, on average, the model's predictions deviate from the actual consumption values by about 2.4 million units. The very high MSE value of 8,314,131,730,197.92 indicates that there are some substantial prediction errors, where the model's forecasts are significantly off from the actual values. An RMSE of approximately 2.88 million units of consumption indicates substantial prediction errors. An R^2 value of 0.18 means that the LSTM model explains only 18% of the variability in the consumption data. This low R^2 score suggests that the model does not capture much of the underlying patterns in the data. A MAPE of 3.33% indicates that, on average, the model's predictions are within 3.33% of the actual consumption values. This relatively low percentage error suggests that while the model makes some large errors (as indicated by the MSE and RMSE), it is generally accurate within a small margin for most predictions. The model shows moderate performance in terms of prediction accuracy. The MAE and MAPE values indicate that the model performs reasonably well on average, with an error margin of	 The graph comparing actual and predicted consumption values further illustrates these points, showing that the model tends to smooth out the fluctuations in actual consumption, leading to underestimation of peaks and overestimation of troughs. This behavior suggests that while the LSTM model captures the general trend, it struggles with capturing the full extent of variability in the data.

	around 3.33%. However, the high MSE and RMSE values reveal the presence of significant prediction errors in some cases, suggesting that the model might not handle all scenarios equally well. The low R^2 score further indicates that the model explains only a small portion of the variability in the data, highlighting the need for further model refinement or the inclusion of more informative features to improve predictive performance.	
Random Forest Regressor Eval MAE: 1637666.23 MSE: 3825364190633.51 RMSE: 1955853.83 R^2(%): 62.292235 MAPE: 2.275008	The MAE of 1,637,666.23, indicates that on average, the model's predictions deviate from the actual values by approximately 1.64 million units of consumption. The Mean Squared Error (MSE) was 3,825,364,190,633.51, highlighting the average squared difference between predicted and actual values. The Root Mean Squared Error (RMSE), which is the square root of MSE, was 1,955,853.83, suggesting that the typical prediction error is around 1.96 million units of consumption. The R^2 Score was 0.6229, meaning that approximately 62.29% of the variance in the actual consumption data is explained by the model's predictions. This score indicates a relatively good fit, with the model capturing a significant portion of the variability in the data. The Mean Absolute Percentage Error (MAPE) was 2.275%, showing that on average, the model's predictions are within 2.28% of the actual consumption values.	 The model performs well in capturing the general trend and variations in the consumption data, as illustrated by the plot of actual vs. predicted values. The model successfully follows the peaks and troughs of the actual consumption, although there are some minimal deviations, overall, the Random Forest Regressor demonstrates a strong predictive capability for this dataset.

Gradient Boosting Regressor MAE: 1714118.73 MSE: 4906490559119.38 RMSE: 2215059.94 R^2(%): 51.635247 MAPE: 2.416074	The MAE indicates that, on average, the model's predictions are off by about 1.71 million units of consumption. Given the actual consumption values range from around 60 to 80 million units, this represents a relatively small error. The RMSE, being in the same units as the data, indicates a modest prediction error. An RMSE of 2.22 million units is higher than the MAE, indicating that there are some errors although not overwhelmingly significant in the context of the actual consumption values. The MAPE directly indicates that the model's predictions are, on average, within 2.42% of the actual values, indicating that the prediction errors are quite small as a percentage of the actual consumption, reinforcing the idea that the model's overall performance is reasonable. The R^2 score indicates that 51.6% of the variance in the actual consumption data is explained by the model. While this is not very high, it shows that the model captures over half of the variability in the data.	 The graph of actual vs. predicted consumption values shows that the Gradient Boosting Regressor model captures the overall trend quite well, although it misses some of the peaks and troughs. The model tends to smooth out some of the variability, leading to less responsiveness to more significant fluctuations in consumption. But overall, the model provides a reasonably accurate forecast for problem study.
LightGBM Regressor Evaluation: MAE: 1940369.23 MSE: 7416177716797.77 RMSE: 2723266.00 R^2(%): 26.896506 MAPE: 2.730060	The LightGBM Regressor, the MAE is approximately 1.94 million units of consumption. This indicates that, on average, the model's predictions deviate from the actual consumption values by about 1.94 million units. Given the actual consumption values range from around 60 to 80 million units, this represents a moderate error. The MSE of 7,416,177,716,797.77 and the RMSE, being in the same units as the data, indicates the typical magnitude of prediction errors. An RMSE of 2.72 million units is higher than the MAE, represents a modest error, suggesting that while the model makes some errors, they are not overwhelmingly significant in the context of the actual consumption values. An R^2 value of 0.27 means that the LightGBM Regressor explains 27% of the variability in the consumption data. This relatively low R^2 score suggests that the model captures some but	 Looking at the graph of actual vs. predicted consumption values, we can see that the model follows the general trend capturing the peaks but tend to smooth out the fluctuations in actual consumption overestimating some troughs. This suggests that while the LightGBM Regressor captures the overall pattern and trend, it struggles with capturing the full extent of variability in the data and may not be as responsive to more significant fluctuations in consumption.

	not a significant portion of the underlying patterns in the data, indicating less correlation between the dependent and independent variables. A MAPE of 2.73% indicates that, on average, the model's predictions are within 2.73% of the actual consumption values. This relatively low percentage error suggests that the model is generally accurate within a small margin for most predictions.	The LightGBM Regressor demonstrates moderate performance in terms of prediction accuracy. The MAE and MAPE values indicate that the model performs reasonably well on average, with an error margin of around 2.73%. The errors (MAE and RMSE) are moderate when viewed in the context of the actual consumption values, representing approximately 2.77% and 3.89% of typical consumption levels, respectively

4.5.3 Comparative analysis of selected ML models

The evaluation outputs of all the different ML models explored are presented together in Table 4. 4 for easier comparison and selection of the best performing model.

Table 4. 4: All ML Models Evaluation Results

Models Evaluation Me	RMSE	R^2 Score	MAPE
LSTM	2.883e+06	0.18	3.333
Random Forest	1.956e+06	0.623	2.273
Gradient Boost Regressor	2.216e+06	0.516	2.416
Light GBR	2.733e+06	0.269	2.73

Analysis and best performing ML model

RMSE: The Random Forest Regressor has the lowest RMSE (1,955,854), indicating that it has the smallest average error magnitude among the models.

R^2 Score: The Random Forest Regressor also has the highest R^2 score (0.623), suggesting that it explains the most variance in the consumption data.

MAPE: The Random Forest Regressor has the lowest MAPE (2.275%), indicating that it provides the most accurate predictions in terms of percentage error.

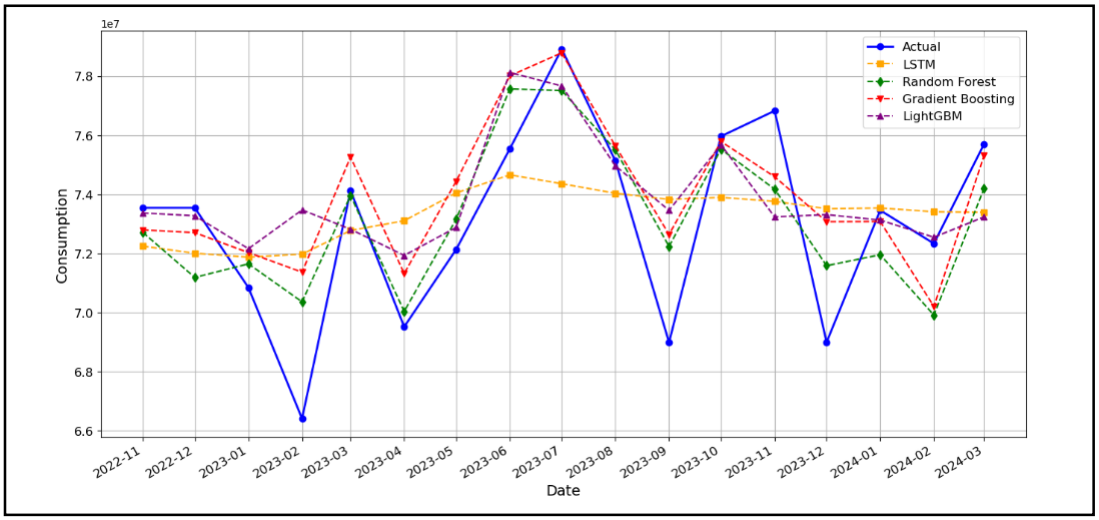


Figure 4. 18: Actual vs predicted electricity demand for all ML models. Source: author

The visual comparison of the models shows that the Random Forest Regressor's predictions are the closest to the actual values, with lower errors and better overall fit. The combined plot of actual vs. predicted values for all models in

Figure 4. 18 visually confirms that the Random Forest Regressor consistently follows the trend of the actual consumption values more closely than the other models.

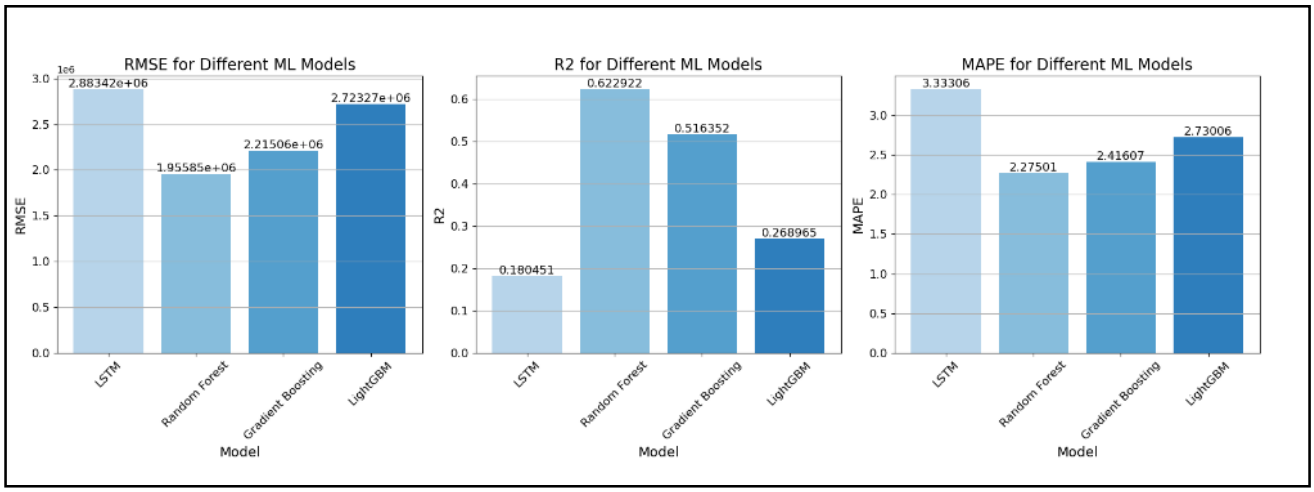


Figure 4. 19: All ML models evaluation results bar plots. Source: author

The bar charts for RMSE, R^2, and MAPE in Figure 4. 19 further highlight the superiority of the Random Forest Regressor in terms of lower error rates and higher explanatory power. The strong performance of the Random Forest Regressor is due to its ability to build multiple decision trees and merges them together to get a more accurate and stable prediction. Random Forest also has the ability to effectively capture complex interactions between features, which is crucial for time series data where historical patterns are important for predicting future values and handles both linear and non-linear relationships well, making it suitable for the mixed nature of time series data with various influencing factors like GDP, population, and renewable energy.

Long Short-Term Memory (LSTM): The LSTM model shows the highest RMSE and MAPE values, indicating that it has the lowest accuracy among the models in terms of predicting the actual consumption values. Its R^2 score is also the lowest, suggesting it has the least explanatory power. The problem involved forecasting based on historical data, which is inherently sequential. LSTM should theoretically perform well due to its design. The poor performance of the LSTM in this case may be due to size of the dataset, which may not be large enough for the model to learn effectively. LSTM is excellent at capturing long-term dependencies, but the electricity consumption can be influenced by various short-term factors such as seasonal changes, economic fluctuations, and population growth, which might overshadow the long-term

patterns. Hence, the advantage of LSTM in capturing long-term dependencies may not be fully realised in this dataset, leading to its relatively poorer performance.

Gradient Boosting Regressor: The Gradient Boosting Regressor performs moderately well, with an RMSE and MAPE that are better than the LSTM but worse than the Random Forest. Its R^2 score suggests it explains a significant portion of the variance but not as much as the Random Forest. The moderate performance of Gradient Boosting is due to its sequential learning process, which allows it to refine predictions iteratively.

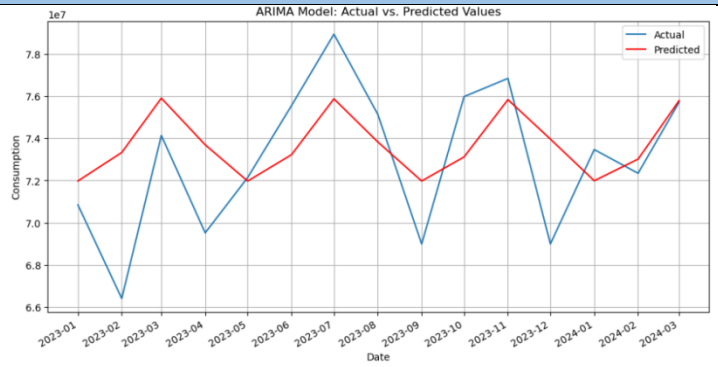
LightGBM Regressor: The LightGBM Regressor has a higher RMSE and MAPE compared to Random Forest and Gradient Boosting but performs better than LSTM. Its R^2 score indicates moderate explanatory power.

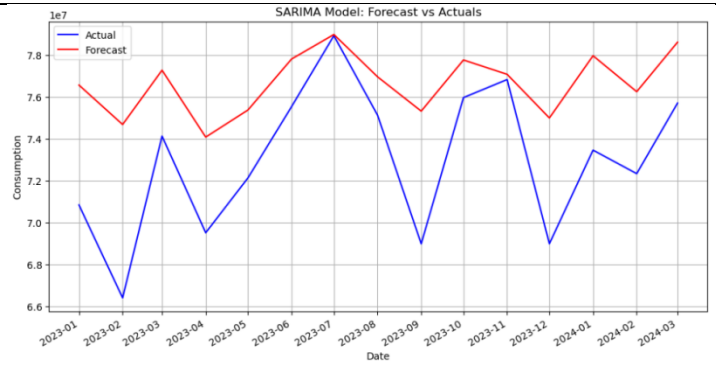
Overall Comparison: Considering all the evaluation metrics, the **Random Forest Regressor** emerges as the best performing model among the models compared, providing the most accurate predictions and the best balance of fit and complexity and making it the most reliable model for forecasting electricity consumption in this context. The Random Forest Regressor's performance in this study aligns with findings from Liaw and Wiener (2020), where they have demonstrated its effectiveness in handling complex datasets with mixed linear and non-linear relationships. They have shown that Random Forest models are particularly well-suited for time series data due to their robustness in managing overfitting and their ability to capture intricate interactions between variables. This robustness is crucial in scenarios involving diverse influencing factors such as GDP, population growth, and renewable energy adoption, as seen in this study. The **Gradient Boosting Regressor** follows as the second best, offering a good compromise between accuracy and complexity. The **LightGBM Regressor** and **LSTM** models, while useful, do not perform as well in this specific application.

4.5.4 Statistical models evaluation analysis

The output results of the statistical models evaluation are presented in Table 4. 5 below with the line plot of the Actual vs predicted values for the test set.

Table 4. 5: Statistical models evaluation analysis

Evaluation Metrics Results	Results Analysis	Actual vs Predicted graph
ARIMA Evaluation: MAE: 2324372.65 MSE: 8759798242373.96 RMSE: 2959695.63 R^2(%): 23.572102 MAPE: 3.261907	The MAE of 2,324,372.65 indicates that, on average, the ARIMA model's predictions deviate from the actual consumption values by approximately 2.32 million units. The RMSE of 2,959,695.63 provides a measure of the standard deviation of the prediction errors. It indicates that the typical prediction error is around 2.96 million units of consumption. The R^2 score of 0.236 suggests that the ARIMA model explains approximately 23.57% of the variance in the actual consumption data. This is relatively low, indicating that the model does not capture a significant portion of the variability in the consumption data. The MAPE of 3.26% indicates that the average absolute error is about 3.26% of the actual consumption values. This is a relative measure, providing an understanding of the error magnitude in relation to the actual values.	 The evaluation metrics show that the ARIMA model has a substantial average error (MAE and RMSE) when predicting consumption values, and it only explains a limited portion of the variance in the data (R^2 score). The relatively low R^2 score suggests that the model might not be capturing all the underlying patterns in the consumption data. However, the MAPE indicates that the errors, while large in absolute terms, represent a modest percentage of the actual consumption values. This suggests that while the model may not be highly precise in predicting specific values, it does reasonably well in maintaining the overall trend, without significant deviations as depicted in the Actual vs Predicted Graph.

SARIMA Evaluation: MAE: 3654140.72 MSE: 18297083155461.23 RMSE: 4277508.99 R^2(%): -59.639248 MAPE: 5.153499	The SARIMA (0, 1, 1)x(1, 0, 1, 12) model, when evaluated on the test data, yields an MAE of 3,654,140.72. This means that, on average, the model's predictions deviate from the actual consumption values by around 3.65 million units. The RMSE stands at 4,277,508.99 indicating that the average deviation of the model's predictions from the actual values is approximately 4.28 million units. The R^2 Score is -0.596392, which is negative and indicates that the model does not capture the variability in the consumption data effectively. The MAPE of 5.153499% indicates that the model's predictions are, on average, about 5.15% off from the actual values. While this percentage error might seem modest, the actual error magnitude (as seen from MAE and RMSE) is substantial due to the large scale of the consumption values.	 The plot comparing actual and forecasted values shows that the SARIMA model's predictions do not align well with the actual consumption values. The forecasted values exhibit a smoother and more stable pattern, while the actual values show more variability and fluctuations. This discrepancy indicates that the SARIMA model fails to capture the peaks and troughs present in the actual consumption data, leading to prediction errors. Overall, the SARIMA model, despite capturing some seasonal patterns, does not perform well in predicting the actual consumption values, as evidenced by the poor visual alignment between the forecasted and actual values and the high error metrics.

4.5.5 Statistical models comparisons

The evaluation outputs of the ARIMA and SARIMA models are presented together in Table 4. 6 for easier comparison and selection of the best-performing model.

Table 4. 6: Statistical models evaluation results

Models Evaluation	RMSE	R^2 Score	MAPE
ARIMA	2.959e+06	0.236	3.262
SARIMA	4.278e+06	-0.596	5.153

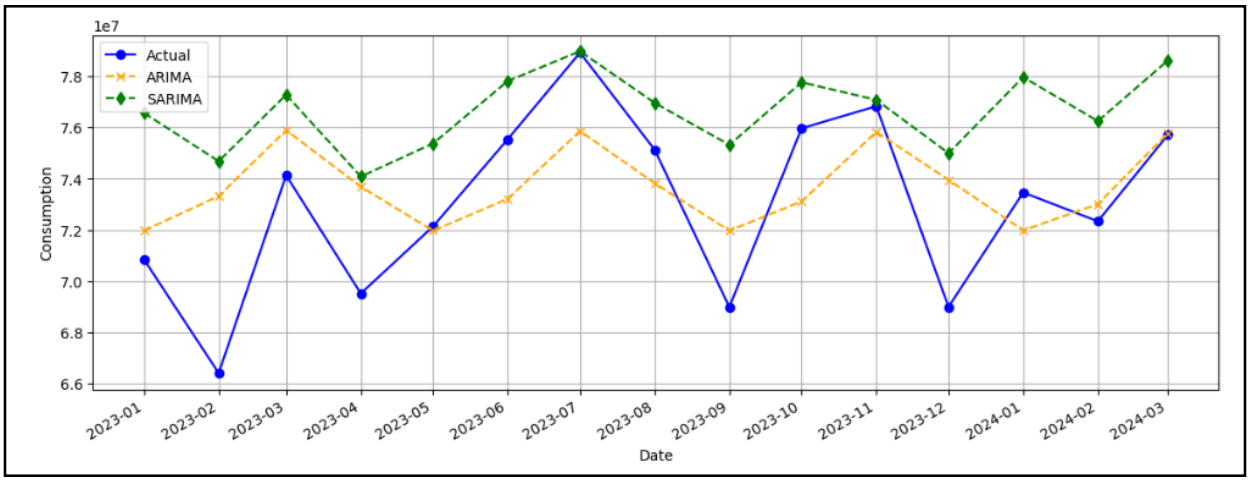


Figure 4. 20: Actual vs predicted line plot for statistical models. Source: author

The visual comparison of the models shows that the ARIMA model's predictions are closer to the actual values, with lower errors and better overall fit. The combined plot of actual vs. predicted values in

Figure 4. 20 for both models visually confirms that the ARIMA model consistently follows the trend of the actual consumption values more closely than the SARIMA model.

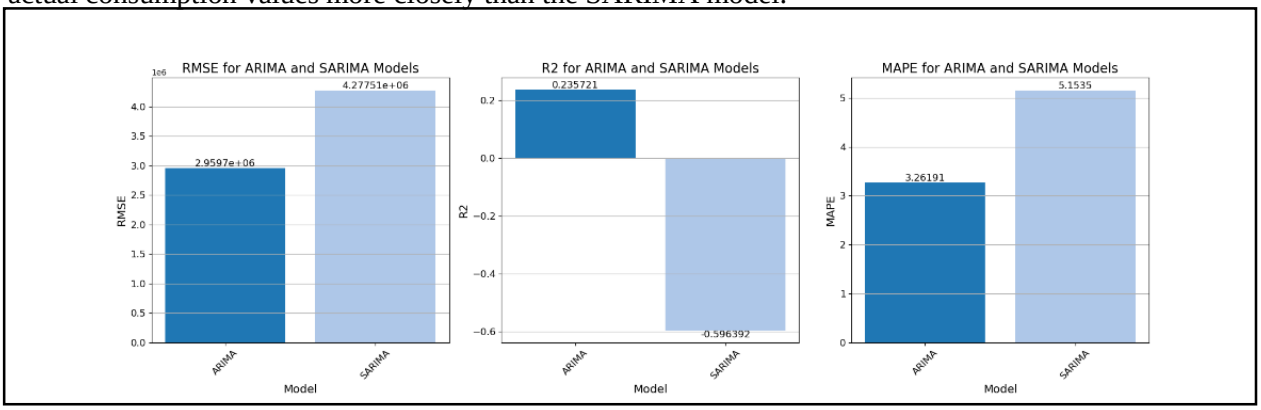


Figure 4. 21: Statistical models evaluation results bar plots. Source: author

The bar charts for RMSE, R^2, and MAPE in Figure 4. 21 further highlight the superiority of the ARIMA model in terms of lower error rates and higher explanatory power.

The ARIMA model, which accounts for non-seasonal trends and patterns, fits the data better under the given circumstances. It is simpler and more effective at capturing the underlying trends without overcomplicating the model. The ARIMA model's strong performance in this study, particularly in capturing non-seasonal trends, is consistent with the observations made by Hyndman and Athanasopoulos (2018), who found that the ARIMA models excel in situations where the data exhibits stable patterns without pronounced seasonality.

The SARIMA model on the other hand, introduces the seasonality components which did not align well with the actual data characteristics. The data did not exhibit strong or regular seasonality, which is why the non-seasonal ARIMA model outperforms the seasonal SARIMA model.

4.5.6 Hybrid model evaluation analysis

The top two performing models from both the Machine Learning and Statistical categories were hybridised to leverage their combined strengths for the forecasting task. The selected models were the Random Forest Regressor from the Machine Learning category and ARIMA from the Statistical models. Two meta-learning techniques, Linear Regression and Gradient Boosting Regression, were employed for the hybridisation. The technique that produced the most accurate results was ultimately chosen for the energy forecasting task.

The evaluation outputs of the hybrid models (Linear Regression and Gradient Boosting Regression) and the individual best performing models (Random Forest and ARIMA) are presented together for easier comparison and selection of the best-performing hybrid model.

Table 4. 7: Hybrid model evaluation results

Models Evaluation	RMSE	R^2 Score	MAPE
LR Hybrid	2.8119e+06	0.3101	3.2289
GBR Hybrid	1.6196e+06	0.9999	0.0199
Random Forest	1.956e+06	0.623	2.273
ARIMA	2.959e+06	0.236	3.262

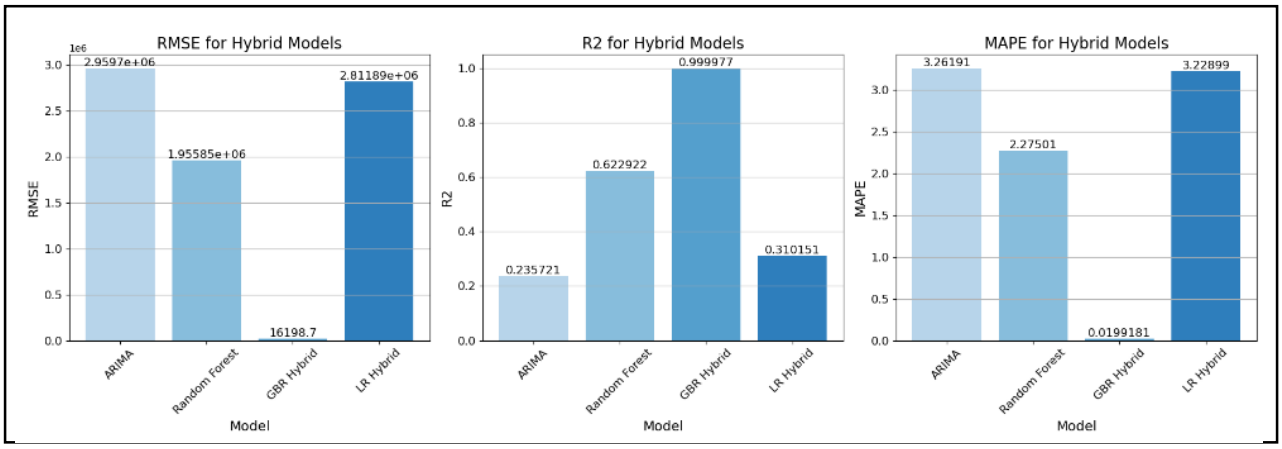


Figure 4. 22: Hybrid models evaluation results bar plots. Source: author

As depicted in Table 4. 7 and Figure 4.22, the LR hybrid demonstrated an RMSE (2,811,888) slightly lower than ARIMA but higher than Random Forest. It achieved an intermediate R^2 score (0.310151), explaining more variance than ARIMA but less than Random Forest and showed a MAPE (3.228988%) slightly better than ARIMA but worse than Random Forest. The GBR Hybrid model on the other hand shows a significant improvement in accuracy compared to the individual models (Random Forest and ARIMA) and the LR hybrid model. The GBR Hybrid model has a drastically lower MAE compared to the other models, indicating that it makes much smaller errors on average. The R2 value for the GBR Hybrid model is nearly perfect, indicating that it explains almost all the variability in the target variable.

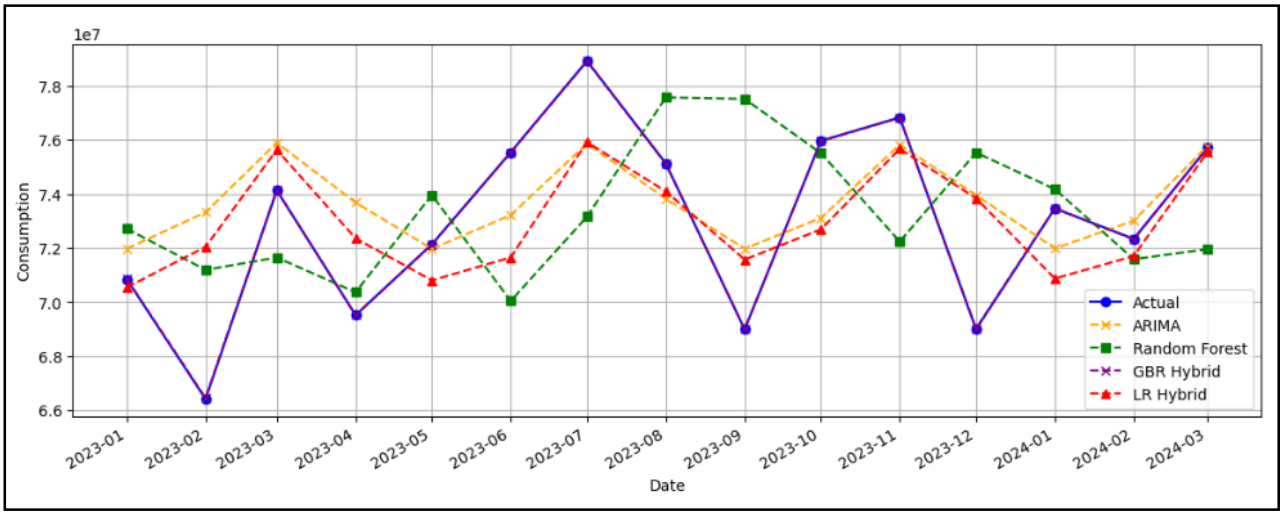


Figure 4. 23: Hybrid models actual vs predicted line plot. Source: author

The plot comparing actual and forecasted values in Figure 4. 23 shows that the LR hybrid model, while showing improvements over ARIMA, did not outperform the Random Forest model. The results suggest

that integrating the strengths of both machine learning and statistical models through hybridisation can yield better results, but the choice of meta-learner and the inherent capabilities of the base models significantly impact the final performance. The GBR hybrid model closely follows the actual values which are represented by blue lines with circular markers, indicating its superior predictive capability. It perfectly captures the peaks in July 2023 and December 2023 more accurately than the other models, thus aligning closely with the actual data. The GBR Hybrid model shows minimal deviation from the actual values, demonstrating its effectiveness in capturing the underlying patterns in the consumption data, making it highly suitable for energy forecasting tasks.

4.5.7 Hybrid models comparison

GBR performed better in combining the results of Random Forest Regression and ARIMA models compared to Linear Regression for several reasons:

Handling non-linearity: GBR is a powerful ensemble method that sequentially builds multiple decision trees, each trained to correct the errors of the previous one, thus allowing GBR to capture complex, non-linear relationships in the data. In contrast, Linear Regression assumes a linear relationship between features and the target variable, which can lead to poor performance when the actual relationship is non-linear.

Bias-variance tradeoff: GBR balances bias and variance effectively by iteratively reducing both training and prediction errors through boosting, leading to better performance on unseen data. In contrast, Linear Regression, despite generally having low variance, can suffer from high bias if the true relationship is not linear, resulting in poor predictive performance.

Interactive effects: GBR can naturally capture interaction effects between features because decision trees can model such interactions. This ability is beneficial when combining outputs from different models, like Random Forest and ARIMA, which may have complex interactions. On the other hand, Linear Regression typically does not capture interaction effects unless explicitly included in the model, which can limit its effectiveness in complex scenarios.

4.5.8 All models evaluations

The evaluation results for all the models explored in this study are presented in Table 4. 8 for comparisons.

Table 4. 8: All models evaluation results

Models Evaluation Metrics	RMSE	R^2 Score	MAPE
ML Before Hyperparameter Tuning			
LSTM	4.452242e+06	-0.955699	4.929366
Random Forest	2.22088e+06	0.513096	2.632919
Gradient Boost Regressor	2.341840e+06	0.45878	2.733379
Light GBR	2.736786e+06	0.260878	2.924546
ML After Hyperparameter Tuning			
LSTM	2.883242e+06	0.180451	3.333058
Random Forest	1.955856e+06	0.622922	2.272501
Gradient Boost Regressor	2.215604e+06	0.516352	2.416077
Light GBR	2.732786e+06	0.268985	2.730046
Traditional Statistical Models			
ARIMA	2.959720e+06	0.235721	3.26191
SARIMA	4.277514e+06	-0.596392	5.15315
Hybrid Models			
LR Hybrid	2.811890e+06	0.310151	3.22899
GBR Hybrid	1.61957e+06	0.999977	0.019918

The evaluation metrics presented in the table indicate that the GBR Hybrid model is the best performing model, with the lowest RMSE, the highest R^2 score, and the lowest MAPE. This demonstrates its high accuracy and reliability in forecasting electricity demand. Conversely, the SARIMA model performed the worst, with the highest RMSE, a negative R^2 score, and the highest MAPE, indicating inaccuracies in its predictions Hyperparameter tuning significantly improves model performance, and the hybrid approach, particularly the GBR Hybrid, offers superior forecasting accuracy compared to standalone machine learning or statistical models.

The superior performance of the GBR Hybrid model, which combines the strengths of Random Forest and ARIMA, mirrors the outcomes of other studies that have employed hybrid approaches for time series forecasting. For example, the work of Ahmed et al. (2020) demonstrated that hybrid models leveraging both statistical and machine learning techniques consistently outperformed standalone models, particularly in complex forecasting environments.

4.6 ELECTRICITY CONSUMPTION FORECAST

The electricity forecast for the next two years (from April 2024 to March 2026) was forecasted using the best performing hybrid model. The GBR hybrid model, by leveraging the strengths of both the Random Forest and ARIMA models, provides a more balanced and potentially more accurate forecast. It captures the necessary seasonality and trends while avoiding overfitting to noise, making it a robust choice for energy demand forecasting over the next two years. The results of this model are captured in

Figure 4. 24 and Figure 4. 25 below.

The Hybrid model, combining the strengths of both Random Forest and ARIMA, shows a more balanced cyclical pattern. It captures the peaks and troughs more distinctly with the forecasted values ranging from approximately 6.97 to 7.79 million.

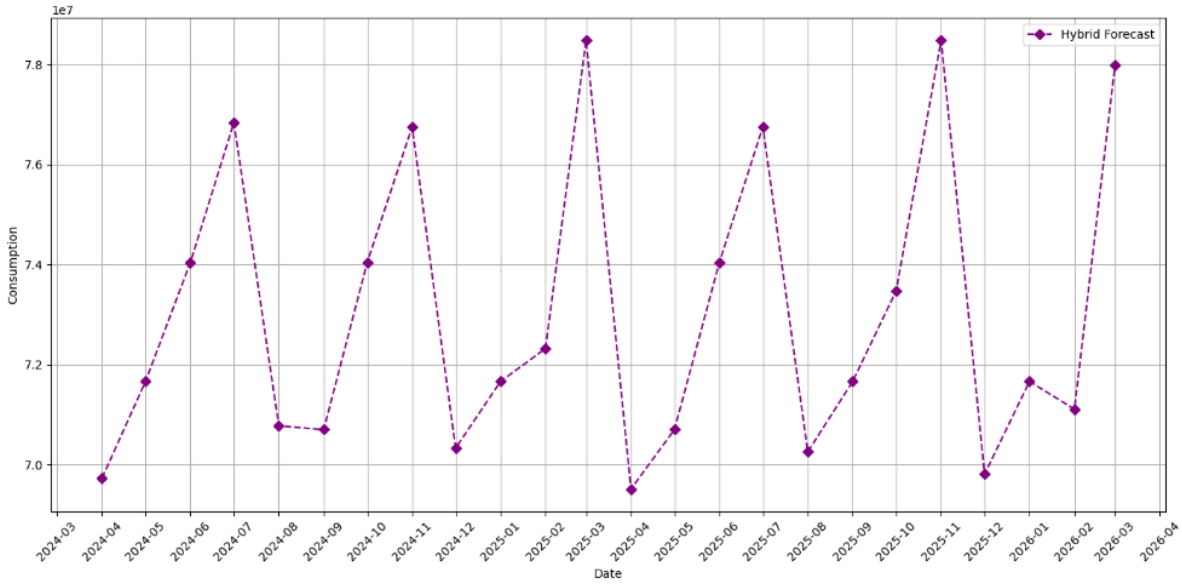


Figure 4. 24: Hybrid model consumption forecast for the next two years. Source: author

	Date	Hybrid_Forecast
0	2024-04-01	6.972736e+07
1	2024-05-01	7.166178e+07
2	2024-06-01	7.404203e+07
3	2024-07-01	7.683906e+07
4	2024-08-01	7.077490e+07
5	2024-09-01	7.069876e+07
6	2024-10-01	7.404203e+07
7	2024-11-01	7.675439e+07
8	2024-12-01	7.032813e+07
9	2025-01-01	7.166178e+07
10	2025-02-01	7.231992e+07
11	2025-03-01	7.848093e+07
12	2025-04-01	6.951214e+07
13	2025-05-01	7.069876e+07
14	2025-06-01	7.404203e+07
15	2025-07-01	7.675439e+07
16	2025-08-01	7.025401e+07
17	2025-09-01	7.166178e+07
18	2025-10-01	7.347489e+07
19	2025-11-01	7.848093e+07
20	2025-12-01	6.980724e+07
21	2026-01-01	7.166178e+07
22	2026-02-01	7.110719e+07
23	2026-03-01	7.798708e+07

Figure 4. 25: Electricity consumption forecast numerical values data frame. Source: author

The electricity consumption forecast results from the hybrid model present a detailed projection of energy consumption for the City of Windhoek over the next two years, underscoring the effectiveness of integrating machine learning and statistical approaches to achieve a more accurate and reliable energy demand prediction. The forecasted values fall within a realistic range in comparison to the historical data, indicating that model provides a credible future consumption forecast.

The forecasted values suggest that the monthly electricity demand for the City of Windhoek will experience minor fluctuations but will remain generally stable over the next two years. The demand peaks are expected around mid-year and at the end of the year, due to seasonal and holiday-related factors.

The consistency, seasonal accuracy, and realistic values of the forecasts reflect the model's robustness and its suitability for practical application in energy planning and management. This result, therefore, successfully addresses the core objective of developing a hybrid forecasting model that combine both the

traditional time series methods and machine learning algorithms for electricity demand forecasting for the City of Windhoek.

4.7 MODEL VALIDATION

This section provides the validation of the hybrid model which was carried out by comparing the predicted consumption values from the model with the actual electricity consumption data for the City of Windhoek during the period from April 2024 to August 2024 as shown in Figure 4. 26. Importantly, the data used for this validation was not available at the time of model training and testing. This ensured that the validation process provides a true assessment of the model's ability to generalise to unseen data and predict future electricity demand accurately.

Date	Actual	Predicted
2024-04-01	70044144.00	6.972736e+07
2024-05-01	72428580.04	7.166178e+07
2024-06-01	76510044.02	7.404203e+07
2024-07-01	78939276.10	7.683906e+07
2024-08-01	74089884.07	7.077490e+07

Figure 4. 26: Model validation actual vs predicted electricity consumption. Source: author

Figure 4. 27 below shows a comparison of actual and predicted electricity consumption for this validation period. As seen in the graph, the model effectively follows the overall trend in electricity consumption with minor deviations, particularly in June and August.

A comparison of the model prediction vs actual demand presented in Figure 4.27 consistently exhibits a nominal negative variance. Though both negative and positive variance between NMD and AD are costly for the City of Windhoek, a negative variance is comparatively less costly, as discussed in section 4.4.5

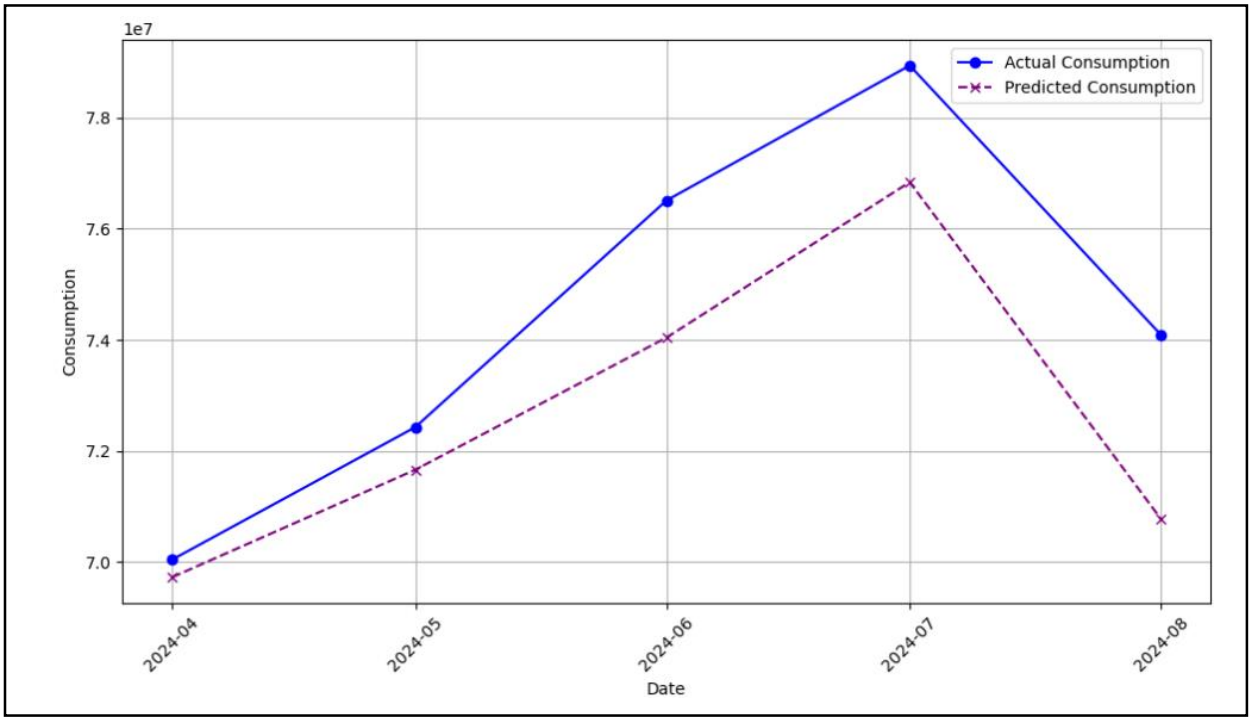


Figure 4. 27: Actual energy consumption (kWh) vs hybrid model forecast. Source: author

```
RMSE: 2106155.0090186363
MAPE: 2.37430379017007%
R²: 0.5380575317466778
```

Figure 4. 28: Hybrid model performance evaluation. Source: author

The model validation results as shown in Figure 4. 28 confirm that the hybrid approach, combining machine learning and statistical methods, is effective for forecasting electricity demand. The relatively low RMSE and MAPE values demonstrate the model's accuracy, while the R^2 score of 53% reflects its reasonable fit to the actual data. The minor discrepancies between actual and predicted values could be attributed to the inherent variability in electricity consumption, influenced by factors such as unexpected weather changes or short-term economic fluctuations.

4.8 SURVEY OUTCOME RESULTS ANALYSIS

This section presents the qualitative findings from the survey conducted among energy experts in the electricity distribution industry. The results are discussed in the context of the historical trends and data

analysed in earlier sections to provide a comprehensive understanding of the key factors impacting the electricity demand in Windhoek (Refer to Appendix B for the survey questions and raw responses).

4.8.1 Demographic overview

A total of 37 experts participated in the survey. Of these, 67% hold engineering positions, and 56% have over ten years of experience in the energy sector, ensuring that the insights provided are both relevant and informed. Furthermore, 58% of the participants are employees of the City of Windhoek, while 14% are from NamPower, which is the national electricity power utility company of Namibia. The profile of respondents marks the reliability of the survey results, as the participants are professionals directly involved in the local electricity distribution and generation sectors. These industry experts provided relevant and informed viewpoints on electricity demand trends, correlation between demand and influencing factors, implications of the variance between NMD and AD, and policy impact on RE adoption in Windhoek.

4.8.2 Trends in electricity demand

Survey respondents were asked to identify trends in electricity demand over the past decade. The majority (47%) indicated that electricity demand has been steadily increasing, while 33% perceived a more rapid increase. However, these perceptions do not fully align with the historical data, which shows significant fluctuations and variability in demand. Although there were periods of increased consumption, particularly between 2016 and 2018, the overall trend reveals notable dips and surges, suggesting that electricity demand has not followed a consistently upward trajectory. This discrepancy highlights the complexity of factors influencing electricity demand in the City of Windhoek.

4.8.3 Correlation between demand and influencing factors

Survey respondents highlighted several key factors influencing electricity demand. A significant majority (64%) agreed that economic growth directly impacts electricity demand. This view aligns with historical data showing that periods of economic expansion often coincide with increased electricity consumption. Respondents noted that the growth in industrial activities and urbanisation are major contributors to this trend, emphasising the role of economic factors in driving electricity demand.

However, the survey also revealed challenges hindering the broader adoption of renewable energy, which could otherwise mitigate some of this demand. Respondents identified high initial investment costs, regulatory barriers, and limited public awareness as the main obstacles to renewable energy adoption. These barriers explain the slower-than-expected growth in renewable energy adoption observed in historical data, particularly before 2018.

Additionally, there was a mixed response regarding the effectiveness of current policies in promoting renewable energy. While some respondents believed that existing policies moderately support renewable energy initiatives, others pointed to significant obstacles and gaps in policy implementation. This division suggests that policy effectiveness may be inconsistent, thus potentially contributing to the uneven pace of renewable energy integration and its impact on electricity demand.

4.8.4 Implications of the variance between NMD and AD

Survey respondents provided valuable insights into the implications of the variance between the NMD and AD for the City of Windhoek. A majority of the experts expressed concern over the financial impact of maintaining a higher NMD than what is typically utilised. They noted that consistently operating below the NMD could result in inefficient resource allocation and unnecessary financial costs for the city, as payments are made based on the contracted maximum demand rather than actual consumption.

Moreover, some respondents highlighted the risks associated with underestimating the NMD. If the actual demand exceeds the NMD, the city could face penalties and higher charges, which could strain the budget. This risk is particularly significant during peak demand periods, where unexpected surges in electricity consumption could lead to financial penalties and supply disruptions.

4.8.5 Survey and historical data

The survey results, combined with historical data, provide a comprehensive understanding of electricity demand trends, key influencing factors, and the impact of renewable energy integration in Windhoek. Survey participants observed a steady or rapid increase in electricity demand over the past decade, consistent with historical peaks around 2016-2018 and the impact of COVID-19 in 2020. Economic growth and industrial activities were identified as primary drivers, reinforcing the correlation between GDP and electricity demand. Additionally, the increase in the adoption of renewable energy post-2018 corresponds with a decline in traditional electricity demand, highlighting renewable energy's growing role in meeting the city's energy needs. However, the historical fluctuations in actual maximum demand, often exceeding the notified maximum demand before 2018, point to inefficiencies, while post-2018, the underutilisation of the increased notified maximum demand led to financial losses, thus mirroring concerns raised by survey respondents about discrepancies between actual and notified demand.

4.9 CHAPTER SUMMARY

This chapter presented the detailed analysis and evaluation of the models used in forecasting electricity consumption, along with the survey outcomes related to electricity demand trends and factors. Table 4. 9 below encapsulates and summerises the key components within the sections of this chapter.

Table 4. 9: Results and discussion of findings summary

Section	Summary
4.1 Chapter Overview	Overview of the chapter's framework, introducing the purpose of the chapter.
4.2 Trends in Electricity Demand	The analysis of electricity demand trends in Windhoek reveals a cyclical pattern with seasonal fluctuations, showing higher consumption during the winter months (June-August) due to increased heating demand. There was an upward trend in energy use between 2014 and 2017, stabilisation from 2017 to 2020, followed by a decline around 2020 due to the COVID-19 pandemic, and a subsequent period of variability.
4.3 Correlation between Demand and Influencing Factors	The analysis shows that electricity demand is positively correlated with GDP trends, with both peaking around 2016-2018 and declining around 2020, likely due to the COVID-19 pandemic. A strong negative correlation exists between electricity demand and both population growth and renewable energy adoption, indicating that as renewable energy usage increases, electricity demand decreases.
4.4 Variance between NMD and AD	The comparison of NMD and AD shows that the AD frequently exceeded the NMD before 2018, indicating inaccurate initial NMD settings. Post-2018, AD remained well below the increased NMD, leading to underutilisation. The financial losses due to underutilised capacity, showed a significant increase in losses post-2018, with an estimated total financial loss of N$ 288,242,177 over the study period, due to over-provisioning.
4.5 Model Performance Evaluation	The Gradient Boosting Hybrid model achieved superior performance with the lowest RMSE (1,619,570), highest R^2 score (0.999977), and lowest MAPE (0.0199%). It demonstrated excellent predictive accuracy compared to other hybrid and standalone models.
4.6 Electricity Consumption Forecast	The GBR electricity consumption forecast for the next two years (April 2024 - March 2026), ranged between 6.97 to 7.79 million kWh. The forecast suggests stable electricity demand with minor fluctuations due to seasonal and holiday factors.
4.7 Model Validation	Model validation performed on actual data (April-August 2024), achieved RMSE of 2.1 million, MAPE of 2.37%, and R^2 of 0.54, based on data not used during training or testing.

4.8 Survey Outcome Results Analysis	**Key Findings from Experts**: Survey participants noted a steady or rapid increase in electricity demand over the past decade. Economic growth and renewable energy integration were identified as critical factors.

CHAPTER FIVE: CONCLUSIONS

5.1 CHAPTER OVERVIEW

This chapter concludes the study by summarizing the overall findings, presenting evidence of the study's deliverables, discussing the challenges encountered, outlining the study's limitations, and highlighting the contributions to the body of knowledge. It also suggests directions for future research. The map shown in Figure 5. 1 below outlines the chapter layout.

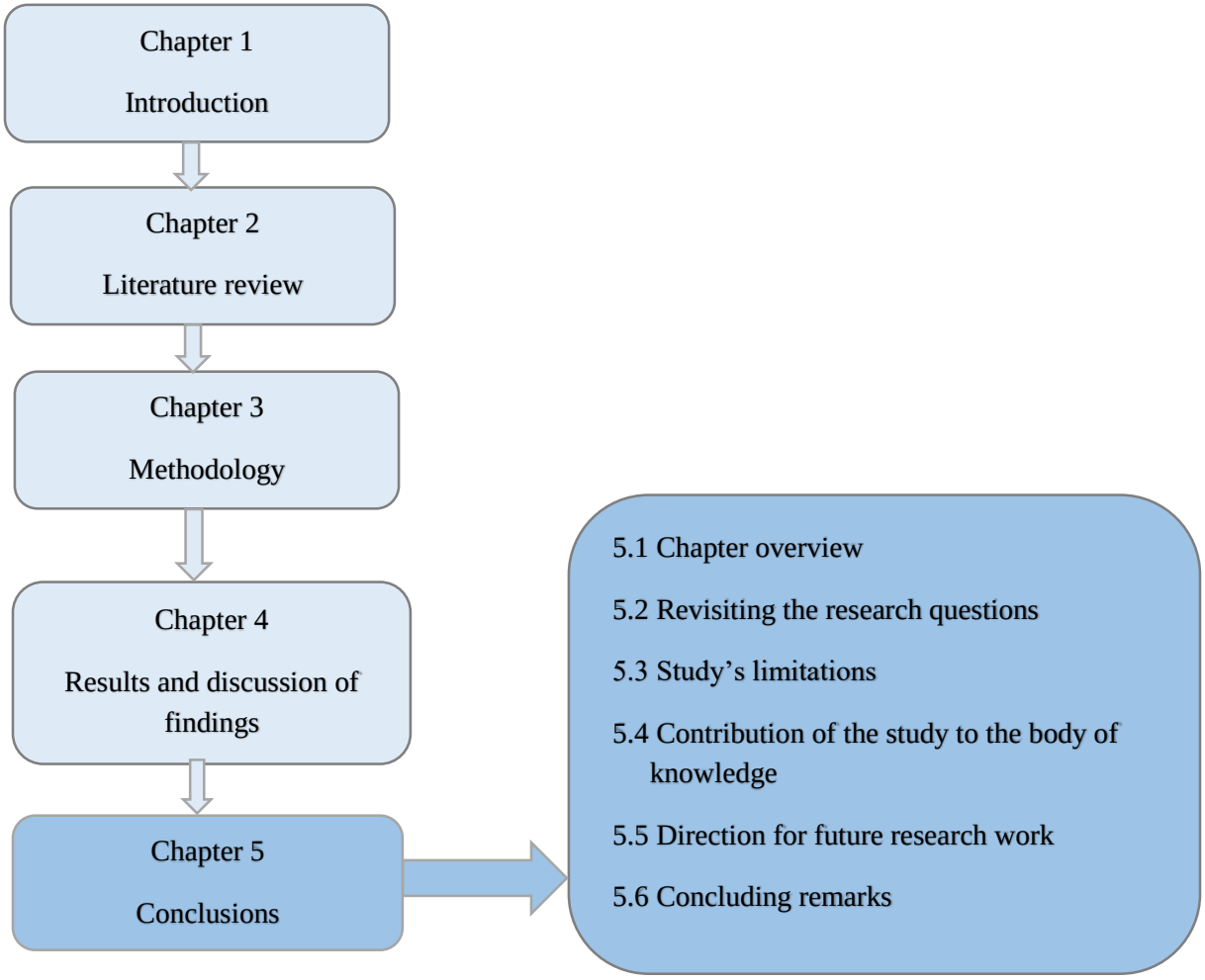


Figure 5. 1: Chapter 5 outline map. Source: author

5.2 REVISITING THE RESEARCH QUESTIONS

This section revisits the research questions outlined at the beginning of the study. The discussion reflects on the accuracy and effectiveness of the hybrid forecasting approach, the identification of key trends and correlations, and the insights gained from both quantitative data and qualitative expert perceptions.

5.2.1 Sub-research question 1

What are the noticeable trends and patterns in electricity demand over the past decade (2014-2023) in Windhoek?

The analysis of historical data focused on the past ten-year period because it captures the most recent trends in electricity consumption, thereby providing a relevant and up-to-date perspective that aligns with current policy and infrastructure planning needs. The analysis revealed that electricity demand in the City of Windhoek shows a cyclical pattern with noticeable fluctuations over the observed period. There are recurrent peaks and dips in the data, indicating seasonal variations affecting electricity consumption with the higher consumption being observed in June, July, and August every year due to heating demands as these are the winter months in Namibia. Over the study period, total annual electricity demand exhibited some fluctuations. The demand increased steadily from 2014 to 2017, stabilised slightly from 2017 to 2019, and then declined between 2020 and 2021. Post-2020, there has been increased variability, with no clear signs of recovery to pre-2020 levels.

5.2.2 Sub-research question 2

What correlations exist between alternative energy adoption/integration, population growth, and GDP?

The study focused on these three factors, namely renewable energy adoption, population growth, and GDP because they are widely recognised as major drivers of electricity consumption. However, the extent of influence on energy consumption, particularly in the context of Windhoek, is not well understood.

Pearson's correlation coefficient was used as the measure of correlation strength. For this study, a correlation coefficient between 0.0 and 0.3 is considered weak, 0.3 to 0.6 is moderate, and values above 0.6 are considered strong. A weak positive correlation was found between GDP and electricity demand, suggesting that while electricity demand tends to increase as GDP rises, GDP does not strongly influence energy consumption. There is a strong negative correlation between Consumption and both Total Cumulative Renewable Energy and Population. This suggests that as renewable energy capacity and

population grow, overall consumption decreases, thus indicating improved energy efficiency and possibly a shift towards sustainable energy practices.

5.2.3 Sub-research question 3

How does the variance between the City of Windhoek's officially Notified Maximum Demand and the observed actual electricity demand influence the city's financial outcomes, particularly in terms of monthly energy procurement?

Significant variances were found between the Notified Maximum Demand and actual demand, leading to financial implications for the city. Aligning the City of Windhoek's notified maximum demand with actual consumption patterns is crucial for optimising financial outcomes, as evidenced by significant fluctuations and inefficiencies before 2018, and underutilisation post-2018. From a qualitative perspective, if the NMD is set too high, the city pays for unused capacity, leading to financial losses due to over-procurement. Conversely, if the NMD is set too low, the city risks incurring penalties for exceeding the notified demand. Therefore, maintaining an NMD that closely matches actual consumption not only avoids these inefficiencies but also ensures more predictable and manageable monthly energy procurement costs.

5.2.4 Sub-research question 4

Which machine learning and statistical model demonstrates the highest accuracy for forecasting electricity demand in Windhoek, and how does the performance of these models compare to the hybrid?

Six (6) models, namely ARIMA, SARIMA, LSTM, Random Forest, (GBR), and LightGBM were explored for forecasting electricity demand for this study. Among these, Random Forest and ARIMA were the top-performing models in their respective categories. However, the hybrid model which combined Random Forest and ARIMA using a stacking ensemble approach, with Gradient Boosting Regressor (GBR) as the meta-learner, demonstrated superior performance in terms of MAPE, RMSE and R^2 score. The comparative analysis highlights the benefits of hybridisation in reducing forecast errors.

5.2.7 Sub-research question 5

What will be the monthly electricity demand for the City of Windhoek for the next two years (April 2024 – March 2026), and how will the demand vary over the months?

The forecast values suggest that the monthly electricity demand for the City of Windhoek will range between 6.97 million kWh and 7.79 million kWh over the next two years, with minor fluctuations but generally stable demand. The demand peaks are expected around mid-year and at the end of the year, due to seasonal and holiday-related factors. For Year 1 (April 2024 - March 2025), the average monthly demand

is estimated at 7.28 million kWh, while for Year 2 (April 2025 - March 2026), the average monthly demand is forecasted to be 7.31 million kWh.

These forecasted values are supported by the model validation results for April-August 2024, which showed strong alignment between actual and predicted demand (refer to Section 4.7), thus providing confidence in the model's accuracy for future predictions.

5.2.6 Sub-research question 6

How do energy experts' perceptions, when analysed alongside historical data, help to identify the key factors influencing the electricity demand?

The analysis of historical data revealed that electricity demand in Windhoek is influenced by several key factors, including economic growth, industrial activities, and renewable energy integration. While the data provided insights into demand fluctuations and correlations with population growth and GDP, certain aspects such as the potential challenges to expanding renewable energy capacity and the effectiveness of energy efficiency initiatives required input from energy experts.

By combining historical data with expert insights, the analysis underscores the importance of strategic investments in power generation and grid improvements. This approach enhances energy security, optimises electricity distribution, and demonstrates the value of integrating expert perspectives with quantitative data for more informed decision-making in managing electricity demand.

5.2.7 Main research question

How can a hybrid forecasting approach, combining traditional time series methods and advanced machine learning techniques, enhance the accuracy and interpretability of electricity demand forecasting for the City of Windhoek?

A hybrid forecasting approach, combining traditional time series methods such as ARIMA with advanced machine learning techniques like Random Forest, enhances the accuracy and interpretability of electricity demand forecasting for the City of Windhoek by leveraging the strengths of both approaches. Traditional time series models, such as ARIMA, excel at capturing linear trends and seasonal patterns, while machine learning models like Random Forest are adept at handling complex, non-linear relationships and high-dimensional data.

The hybrid model integrates these two methods using a stacking ensemble approach, with Gradient Boosting Regressor (GBR) as the meta-learner. This combination allows for more accurate and

comprehensive forecasts, as demonstrated by reductions in performance metrics such as RMSE and MAE, thus showing significant improvements over individual models.

5.3 STUDY'S LIMITATIONS

The primary limitations of this study include its reliance on historical data, which may not fully account for future shifts in electricity consumption due to technological advancements or policy changes. Additionally, the study's sample size was constrained to 123 monthly observations for electricity demand and 40 expert survey responses, which may limit the generalisability of findings. The scope of the study was limited to the City of Windhoek; thus the findings apply in this specific context. To add to that, in the absence of city-level statistics, the study drew inferences from national-level data on population and economic indicators. Furthermore, the sampling technique used in expert surveys followed a purposive sampling approach, meaning that the insights reflect perspectives from selected industry professionals rather than a random sample. Additionally, the assumptions made in the modeling process, such as linear relationships among variables and the absence of multi collinearity, may not always hold true due to the complexity of real-world dynamics. Nonetheless, the study lays an adequate foundation for electricity demand projection and the findings are a good indication of forecasting model performance, which can be extrapolated to other contexts beyond the City of Windhoek.

5.4 CONTRIBUTION OF THE STUDY TO THE BODY OF KNOWLEDGE

The study extends the knowledge on electricity demand forecasting and hybridisation of traditional statistical forecasting methods with advanced machine learning algorithms. Specifically, the study demonstrates the superiority of the hybrid model over individual models in terms of accuracy, thereby offering a more robust tool for forecasting electricity demand. By incorporating economic indicators, population changes, and renewable energy integration into the forecasting model, the research provides a broader approach that addresses the limitations of previous models. The findings contribute to the literature by highlighting the importance of combining traditional and machine learning approaches to improve forecasting outcomes in urban energy management. Moreover, the study offers practical insights which are useful for policymakers and energy distribution analysts.

5.5 DIRECTIONS FOR FUTURE RESEARCH WORK

In light of the aforementioned limitations of the study, directions for future research include the following:

- Investigate the application of hybrid electricity forecasting models in different cities or regions with varying consumption patterns to validate their effectiveness across diverse contexts.
- Develop real-time electricity demand forecasting models by incorporating Internet of Things (IoT) devices and smart meter data to improve the responsiveness and accuracy of predictions.
- Explore the use of hybrid models for forecasting electricity demand across different sectors such as residential, commercial, and industrial, to better tailor energy management strategies.
- Assess the impact of incorporating real-time weather data into hybrid models to enhance predictive accuracy, particularly during extreme weather events.
- Although this current study integrated renewable energy data into the demand forecasting model, further work is needed to enhance the granularity and accuracy of this integration. Future research could explore more sophisticated methods for forecasting renewable energy output and more dynamically incorporate these predictions into electricity demand forecasts.
- Investigate practical deployment strategies for hybrid forecasting models, including cloud-based systems, smart meter integration, and real-time data pipelines to enhance their applicability in energy management.

5.6 CONCLUDING REMARKS

This study embarked on a journey to improve the accuracy of electricity demand forecasting for the City of Windhoek through the development of a hybrid model combining both traditional time series methods and machine learning techniques. By leveraging the strengths of ARIMA for handling non-seasonal components and Random Forest for capturing complex interactions between variables, the research has demonstrated the effectiveness of integrating these methods. The hybrid GBR model outperformed all individual models across key metrics, showcasing the power of combining approaches to address the complexity inherent in electricity demand forecasting.

The findings of this research not only provided accurate electricity consumption forecasts for the next two years but also offered valuable insights into the relationships between electricity demand, economic growth, population trends, and renewable energy integration. These insights are crucial for policymakers and energy planners aiming to ensure a balance between electricity supply and demand, thereby avoiding costly

penalties and optimising resource allocation. The financial implications of discrepancies between the NMD and AD further emphasised the need for dynamic and data-driven approaches to energy management.

As we look to the future, the ever-growing importance of energy sustainability and the integration of renewable energy sources will continue to challenge traditional forecasting methods. This study establishes a foundation for improving the performance of forecasting models through the hybridizing of methods. In an era where energy management is critical to urban development, it is essential to continuously refine our forecasting techniques and embrace innovation. As technology evolves, so too must our methods, ensuring that cities like Windhoek can thrive in an energy-efficient and economically sustainable manner.

In closing, this research highlights the profound impact that accurate electricity demand forecasting can have on the city's finances, operational efficiency, and long-term sustainability. The future of energy management lies not just in our ability to predict but in our capacity to adapt and optimise, thus ensuring that we meet the demands of tomorrow with foresight, precision, and responsibility. The question is not whether we can predict the future, but whether we are prepared to act on what we know.

REFERENCES

Abeywickrama, S., & Dinesh Asanka, P. P. G. (2023). Integrating weather patterns into machine learning models for improved electricity demand forecasting in Sri Lanka. *Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2023*. https://doi.org/10.1109/SCSE59836.2023.10215047

Ahmed, N. K., Atiya, A. F., El Gayar, N., & El-Shishiny, H. (2020). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5), 594–621. https://doi.org/10.1080/07474938.2010.481556

Alam, M. S., & Arefifar, S. A. (2019). Energy management in power distribution systems: Review, classification, limitations and challenges. *IEEE Access*, *7*, 92979–93001. https://doi.org/10.1109/ACCESS.2019.2927303

Alooba. (2024). *Model performance metrics: Everything you need to know when assessing model performance metrics*. Alooba. https://www.alooba.com/skills/concepts/data-science/model-performance-metrics/

Azka, M., Wiradinata, S. A., Faisal, M., & Suhartono. (2020). Double seasonal ARIMA for forecasting electricity demand of Kuaro Main Gate in East Kalimantan. *IOP Conference Series: Materials Science and Engineering*, *846*(1). https://doi.org/10.1088/1757-899X/846/1/012064

Bajaj, A. (2023, August 18). *ARIMA & SARIMA: Real-world time series forecasting*. Neptune. Ai. https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide

Banimustafa, A., & Al-Omari, Z. (2022). Trends of electricity consumption in Jordan. *1st International Engineering Conference on Electrical, Energy, and Artificial Intelligence, EICEEAI 2022*. https://doi.org/10.1109/EICEEAI56378.2022.10050498

Bartz, E., Bartz-Beielstein, T., Zaefferer, M., & Mersmann, O. (n.d.). *Hyperparameter tuning for machine and deep learning with R A practical guide*. https://library.oapen.org/bitstream/id/240dc394-c1a4-414e-8465-d9dc8824c7e5/978-981-19-5170-1.pdf

Berman, J. J. (2016). Understanding your data. In *Data Simplification* (pp. 135–187). Research Gate.

Brownlee, J. (2020). *Time series forecasting as supervised learning*. . https://machinelearningmastery.com/time-series-forecasting-supervised-learning/

Brownlee, J. (2023). *How to create an ARIMA model for time series forecasting in Python*. Machine Learning Mastery. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

Cetinkaya, M., & Acarman, T. (2021). Next-Day electricity demand forecasting using regression∗. *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, 1549–1554. https://doi.org/10.1109/ICAIS50930.2021.9395926

Clark, R. S., & Clark, V. L. P. (2022). The use of mixed methods to advance positive psychology: A methodological review. *International Journal of Wellbeing*, *12*(3), 35–55. https://doi.org/10.5502/ijw.v12i3.2017

Di Persio, L., & Fraccarolo, N. (2023). Energy consumption forecasts by gradient boosting regression trees. *Mathematics*, *11*(5). https://doi.org/10.3390/math11051068

Dixie Electric Power Association. (2024). *Understanding demand*. Dixie Electric Power Association. https://www.dixieepa.com/faqs_category/understanding-demand/

Elsaraiti, M., Ali, G., Musbah, H., Merabet, A., & Little, T. (2021). Time series analysis of electricity consumption forecasting using ARIMA model. *IEEE Green Technologies Conference*, *2021-April*, 259–262. https://doi.org/10.1109/GreenTech48523.2021.00049

Encode. (2023, June 13). *Training, validation, test Split for machine learning datasets*. Encode Blog. https://encord.com/blog/train-val-test-split/

Energy Agency, I. (2024). *Electricity 2024 - Analysis and forecast to 2026*. www.iea.org

Express Analytics. (2021, November 9). *Everything you Need to know about time series analysis – A Primer*. American Express. https://www.expressanalytics.com/blog/time-series-analysis/

GeeksforGeeks. (2024). *Random forest algorithm in machine learning*. GeeksforGeeks.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media Inc.

Giamarelos, N., Papadimitrakis, M., Stogiannos, M., Zois, E. N., Livanos, N. A. I., & Alexandridis, A. (2023). A Machine Learning Model Ensemble for mixed power load forecasting across multiple time horizons. *Sensors*, *23*(12). https://doi.org/10.3390/s23125436

Gibson, K. (2019). Mixed methods research in sport and exercise: Integrating qualitative research. In *Routledge Handbook of Qualitative Research in Sport and Exercise*. Routledge.

Goswami, K., & Kandali, A. (2022, July 1). Electricity demand prediction using data driven forecasting scheme: ARIMA and SARIMA for real-time load data of Assam. *International Conference on Computational Performance Evaluation (ComPE)*.

Graig, A. (2024, June 7). City wants 10.3% electricity hike. *The Namibian Sun*.

Guo, F., Mo, H., Wu, J., Pan, L., Zhou, H., Zhang, Z., Li, L., & Huang, F. (2024). *Citation: A hybrid stacking model for enhanced short-term load forecasting*. https://doi.org/10.3390/electronics

Hasan, M. B., Yousaf, J., Khan, N. T., Aslam, J., & Ehtisham, C. M. (2021). Decision support system for forecasting the trends in energy sector of Pakistan: Multivariate modeling with socio-economic and environmental factors. *2021 International Conference on Computing, Electronic and Electrical Engineering, ICE Cube 2021 - Proceedings*. https://doi.org/10.1109/ICECube53880.2021.9628303

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14), 5481–5487). Copernicus GmbH. https://doi.org/10.5194/gmd-15-5481-2022

Hou, L., Shen, X., & Zhang, L. (2023). Urban electricity demand forecasting with a hybrid machine learning model. *ICNSC 2023 - 20th IEEE International Conference on Networking, Sensing and Control*. https://doi.org/10.1109/ICNSC58704.2023.10319027

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. https://otexts.com/fpp2/

Iweh, C. D., Gyamfi, S., Tanyi, E., & Effah-Donyina, E. (2021). Distributed generation and renewable energy integration into the grid: Prerequisites, push factors, practical options, issues and merits. *Energies*, *14*(17). https://doi.org/10.3390/en14175375

Jaisumroum, N., & Teeravaraprug, J. (2017). *ICIEA 2017 : 2017 4th International Conference on Industrial Engineering and Applications : April 21-23, 2017, Nagoya, Japan*. IEEE Press.

Jie, B., Baba, J., & Kumada, A. (2023). Mathematical analysis of time-series & meteorological factors in electricity demand forecasting based on carbon neutrality. *2023 IEEE IAS Industrial and Commercial Power System Asia, I and CPS Asia 2023*, 137–142. https://doi.org/10.1109/ICPSAsia58343.2023.10294862

Khatoon, S., Nasiruddin, I., Singh, A., & Gupta, P. (2020, December). Effects of various factors on electric load forecasting: An overview. *6th IEEE Power India International Conference (PIICON)*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9921606/

Klyuev, R. V., Morgoev, I. D., Morgoeva, A. D., Gavrina, O. A., Martyushev, N. V., Efremenkov, E. A., & Mengxu, Q. (2022). Methods of forecasting electric energy consumption: A literature review. *Energies*, *15*(23). MDPI. https://doi.org/10.3390/en15238919

Kolawole, A., Brookes, O., & School, B. (2017). *An empirical analysis of energy demand in Sub-Saharan Africa*.https://radar.brookes.ac.uk/radar/items/1886b97b-4870-430a-9815-e0a2cf3ca658/1/

Kpomonè, A. B. K., Kpomonè, A. B. K., Kpomonè, A. B. K., Gnadi, P. E. T., Gnadi, P. E. T., Gnadi, P. E. T., Yao, B., Yao, B., Yao, B., Dosseh, K., & Komla, N. (2024). Multiple linear regression to predict electrical energy consumption based on meteorological data: Application to some sites supplied by the CEB in Togo. *American Journal of Applied Sciences*, *21*(1), 15–27. https://doi.org/10.3844/ajassp.2024.15.27

Krishnan, M., Jung, Y. M., & Yun, S. (2020, March 1). Prediction of energy demand in smart grid using hybrid approach. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. https://ouci.dntb.gov.ua/en/works/963qXoJ7/

Lanka, V. V. S., Roy, M., Suman, S., & Prajapati, S. (2021, February 5). Renewable energy and demand forecasting in an integrated smart grid. *2021 Innovations in Energy Management and Renewable Resources, IEMRE 2021*. https://doi.org/10.1109/IEMRE52042.2021.9386524

Liaw, A., & Wiener, M. (2020). *Classification and regression by RandomForest*. https://www.researchgate.net/publication/228451484

Lindemann, B., Müller, T., Vietz, H., Jazdi, N., & Weyrich, M. (2020). *A survey on long short-term memory networks for time series prediction Industry 4.0 in forming technologies view project verification of distributed automation systems based on a model composition View project ScienceDirect: A survey on long short-term memory networks for time series prediction-review under responsibility of the scientific committee of the 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering*. https://doi.org/10.13140/RG.2.2.36761.65129/1

Liu, L., Yang, J., & Chen, W. (2022). Electricity consumption forecasting in jiangsu province based on machine learning model. *2022 4th International Academic Exchange Conference on Science and Technology Innovation, IAECST 2022*, 613–616. https://doi.org/10.1109/IAECST57965.2022.10061984

Macrotrends. (2024, August 21). *Namibia GDP Per Capita 1960-2024*. Macrotrends. https://www.macrotrends.net/global-metrics/countries/NAM/namibia/gdp-per-capita

Melnikovas, A. (2018). Towards an explicit research methodology: Adapting research onion model for futures studies. *Journal of Futures Studies*, *23*(2), 29–44. https://doi.org/10.6531/JFS.201812_23(2).0003

Mir, A. A., Alghassab, M., Ullah, K., Khan, Z. A., Lu, Y., & Imran, M. (2020). A review of electricity demand forecasting in low and middle income countries: The demand determinants and horizons. In *Sustainability (Switzerland)*, *12*(15). MDPI. https://doi.org/10.3390/SU12155931

Muaz, M. H., Akhtaruzzaman, & Rahman, M. (2024, May 1). Forecasting electricity consumption in five regions of Dhaka City: A comparative assessment among various ML-Boosting odels. *6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*.

Muhammad, S., & Kabir, S. (2018). *Methods of data collection*. https://www.researchgate.net/publication/325846997

Nampower. (2020). *NAMPOWER standard conditions for the supply of electricity revision May 2000 Namibia Power Corporation (PTY) LTD Annexure A Standard Conditions for the supply of electricity*. https://www.nampower.com.na/public/docs/ConditionsOfElectricitySupply.pdf

Organization for Autism Research. (2020, March 9). *Five basic types of research studies*. Organization for Autism Research. https://researchautism.org/blog/five-basic-types-of-research-studies/

Petrović-Ranđelović, M., Kocić, N., & Stojanović-Ranđelović, B. (2020). The importance of renewable energy sources for sustainable development. *Economics of Sustainable Development*. https://scindeks-clanci.ceon.rs/data/pdf/2560-421X/2020/2560-421X2002015P.pdf

Prabhakaran, S. (2023, July 12). *Augmented Dickey Fuller Test (ADF Test) – Must read guide*. Machine Learning Plus.

Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLoS ONE, 18*(2 February). https://doi.org/10.1371/journal.pone.0279774

Román-Portabales, A., López-Nores, M., & Pazos-Arias, J. J. (2021). Systematic review of electricity demand forecast using ann-based machine learning algorithms. *Sensors*, 21(13). MDPI AG. https://doi.org/10.3390/s21134544

Saunders, M. N. (2009). *Understanding research philosophies and approaches*. https://www.researchgate.net/publication/309102603

Shearer, C. (2000). the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin. *Journal of Data Warehousing*.

Shiwakoti, R. K., Charoenlarpnopparut, C., & Chapagain, K. (2023). Time series analysis of electricity demand forecasting using seasonal ARIMA and an exponential smoothing model. *2023 International Conference on Power and Renewable Energy Engineering, PREE 2023*, 131–137. https://doi.org/10.1109/PREE57903.2023.10370319

Skaf, Z., Abdulmouti, H., Alshareif, A., & Albanna, M. H. (2023). Electricity consumption forecasting based on machine learning techniques. *2023 Advances in Science and Engineering Technology International Conferences, ASET 2023*. https://doi.org/10.1109/ASET56582.2023.10180605

Straight, J. (2023, November 28). *What is electricity demand?* Santee Cooper. https://www.santeecooper.com/global-news/2023/112823-Electricity-Demand.aspx

Teacha. (2023, February 1). *Grade 9 Energy generation and the National energy grid in PowerPoint.* Teacha. https://www.teacharesources.com/product/grade-9-energy-generation-and-the-national-energy-grid-in-powerpoint/

The Namibian. (2024, June 30). N$100 will now buy fewer than 40 electricity units. *The Namibian*.

Upadhaya, D., Thakur, R., & Singh, N. K. (2019). A systematic review on the methods of short term load forecasting. *2019 2nd International Conference on Power Energy Environment and Intelligent Control, PEEIC 2019*, 6–11. https://doi.org/10.1109/PEEIC47157.2019.8976518

Wang, S., Tang, J., & Liu, H. (2016). Feature Selection. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1–9). Springer. https://doi.org/10.1007/978-1-4899-7502-7_101-1

Wong, A., Unni, P., Henrique, A. L. K. P., Nguyen, T. A., Chiu, C., Khmelevsky, Y., & Mahony, J. (2022). Machine learning models application in daily forecasting of hourly electricity usage. *SysCon 2022 - 16th Annual IEEE International Systems Conference, Proceedings*. https://doi.org/10.1109/SysCon53536.2022.9773835

Yıldız Doğan, G., Aksoy, A., & Öztürk, N. (2024). A hybrid deep learning model to estimate the future electricity demand of sustainable cities. *Sustainability (Switzerland)*, *16*(15). https://doi.org/10.3390/su16156503

APPENDIX A: ETHICAL CLEARANCE CERTIFICATE

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

FACULTY RESEARCH ETHICS COMMITTEE (F-REC)

DECISION/FEEDBACK ON THE RESEARCH PROPOSAL

Dear Ms Ndakalako Diana Nambahu

RESEARCH TOPIC: HYBRID APPROACH FOR ELECTRICITY DEMAND FORECASTING FOR THE CITY OF WINDHOEK.

Supervisor (if applicable): **Dr Richard Maliwatu**

Qualification registered for (if applicable): Master of Data Science

(Reference number of applications: **FACULTY RESEARCH ETHICS COMMITTEE REGISTRATION NUMBER: FREC - 22/24)**

Re: Ethical screening application No: **FREC - 22/24**

The Faculty of **Computing and Informatics** Ethics Screening Committee of the Namibia University of Science and Technology reviewed your application for the above-mentioned research. The research as set out in the application has been:

Approved	X

(Indicate with an X, and N/A if not applicable and proceed)

We would like to point out that you, as a researcher, are obliged to maintain the ethical integrity of your research, adhere to the ethical guidelines of NUST, and remain within the scope of your research proposal and supporting evidence as submitted to the F-REC. Should any aspect of your research change from the information as presented to the F-REC, which could affect the possibility of harm to any research subject, you are under the obligation to report it immediately to your supervisor or F-REC as applicable in writing. Should there be any uncertainty in this regard, you must consult with the F-REC.

We wish you success with your research and trust that it will make a positive contribution to the questfor knowledge at NUST.

Any ethical issues that need to be highlighted?	Why are these issues important?	What must/could be done to minimize the ethical risk?
No	N/A	N/A

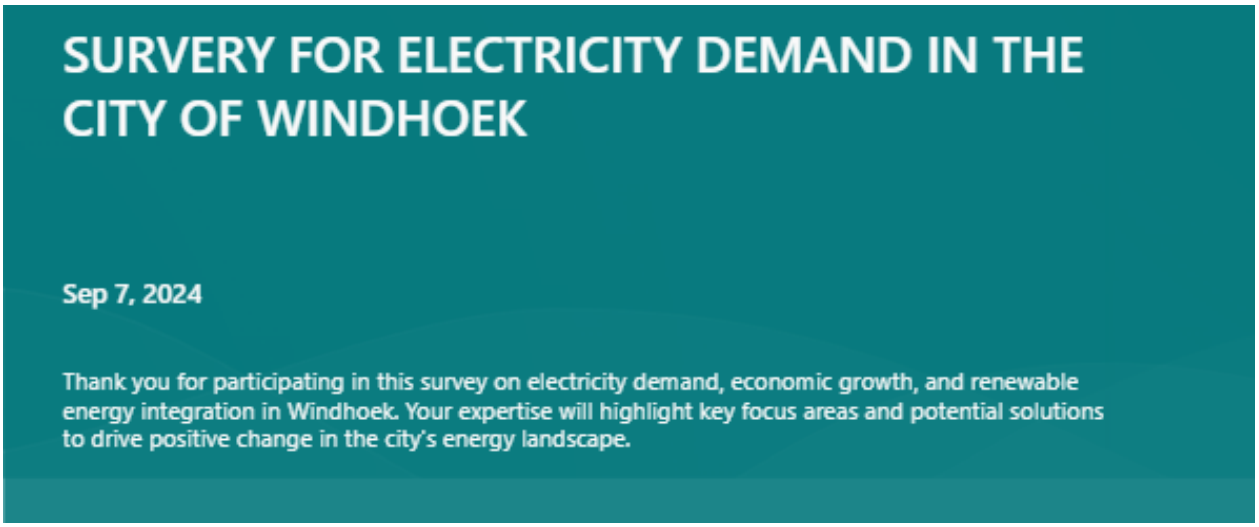
Recommendation: The application is approved.

Sincerely,

Prof. Suama L Hamunyela
Chairperson: Faculty Ethics Screening
CommitteeTel: +264-61-207-2922
CC: Co-supervisor:

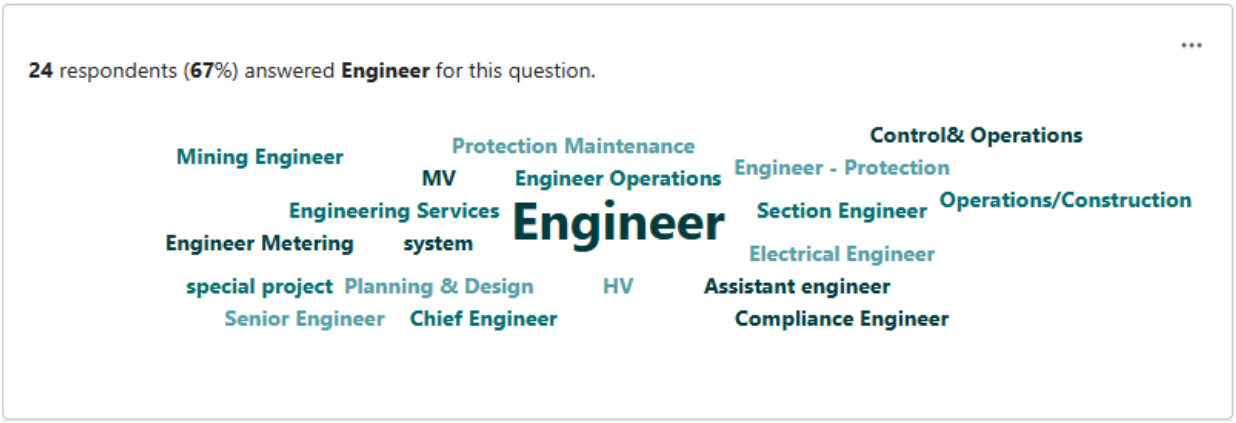
NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY
Pvt Bag 13388 Windhoek NAMIBIA
2024 -05- 15
OFFICE OF THE DEAN
FACULTY OF COMPUTING & INFORMATICS

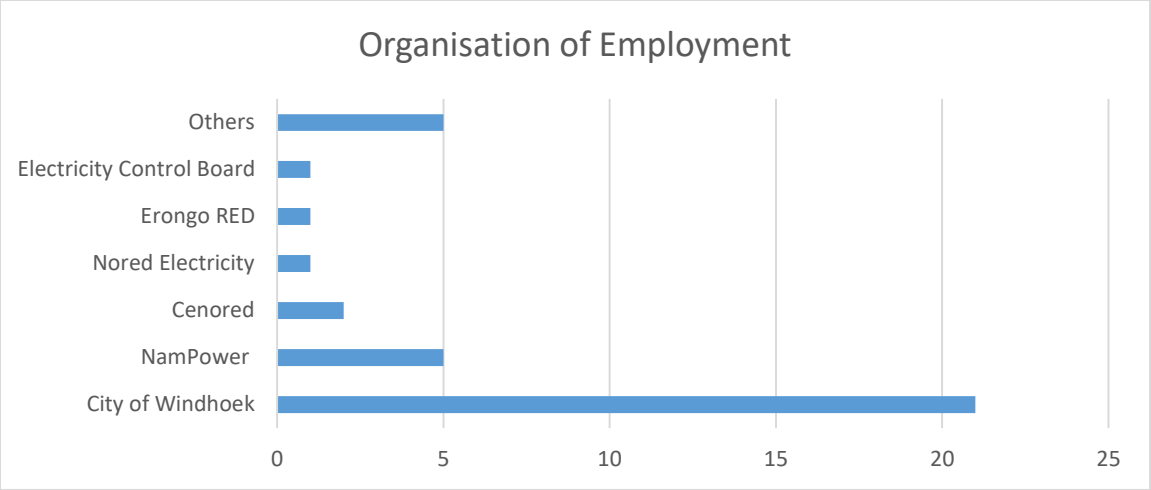
APPENDIX B: SURVEY QUESTIONS AND RESULTS



The survey questions and graphical results presentation of the survey conducted to gather insights from energy experts in the electricity distribution industry are presented in this Appendix. To understand the demographics of the participants, they were asked to provide their job titles, the organizations they work for, and their years of experience in the energy sector. Figures presented in this Appendix were extracted from the Microsoft Forms platform, which was used to gather the survey results.

Participants job titles





3. Years of experience in the energy sector:

More Details

0-2 years	3
3-5 years	6
6-10 years	7
over 10 years	20



4. How would you describe the trends in electricity demand for the City of Windhoek for the past ten years?

More Details

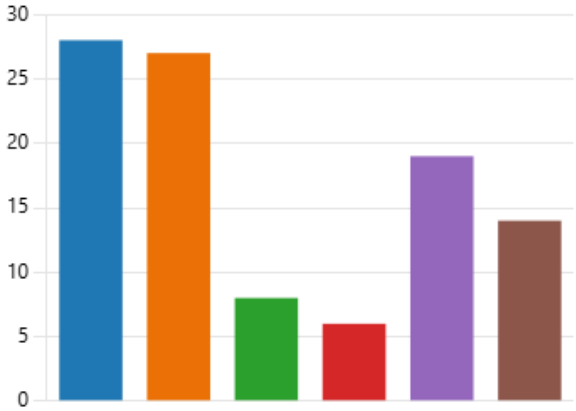
Increasing steadly	17
Increasing rapidly	12
Stable	5
Decreasing	1
Fluctuating	1



5. What factors do you think influence changes in electricity demand?

More Details

Economic growth	28
Population growth	27
Weather conditions	8
Technological advancements	6
Industrial activities	19
Renewable energy adoption	14



6. How does economic growth affect the electricity demand in the City of Windhoek, in your opinion?

More Details

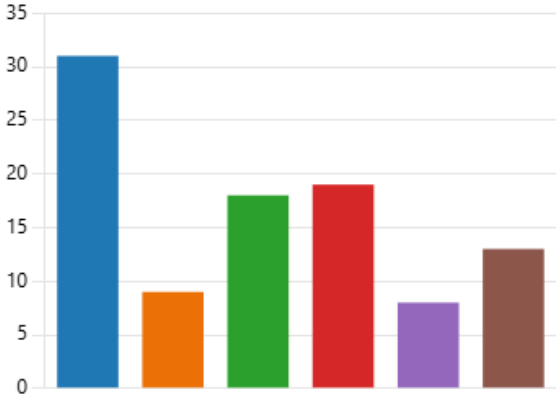
Directly increases demand	23
Moderately increases demand	10
No significant impact	1
Moderately decreases demand	2
Directly decreases demand	0



7. What are the main challenges hindering the widespread adoption of renewable energy in the City of Windhoek?

More Details

High initial investment costs	31
Lack of infrastructure	9
Regulatory barriers	18
Limited public awareness	19
Technological limitations	8
Uncertainty in energy supply	13



8. How do existing policies and regulations support or hinder renewable energy adoption in the City of Windhoek?

More Details

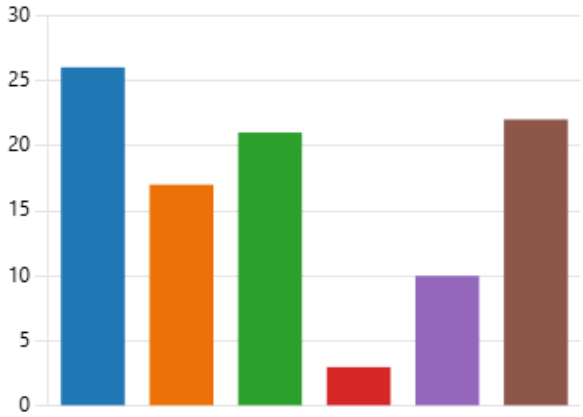
Strongly support	11
Somewhat support	8
Neutral	5
Somewhat hinder	11
Strongly hinde	1



9. What are the key considerations policymakers should prioritize when drafting energy-related policies for electricity distributors like City of Windhoek?

More Details

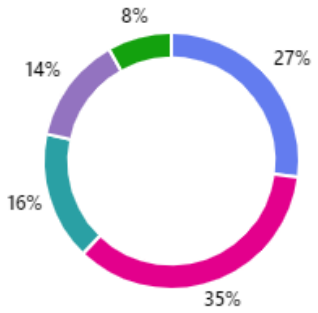
Ensuring energy security	26
Promoting renewable energy	17
Balancing costs and benefits	21
Addressing climate change	3
Encouraging private sector inve...	10
Improving grid infrastructure	22



10. How does the actual electricity demand in Windhoek compare to the officially notified maximum demand, and how does this difference affect the city's finances?

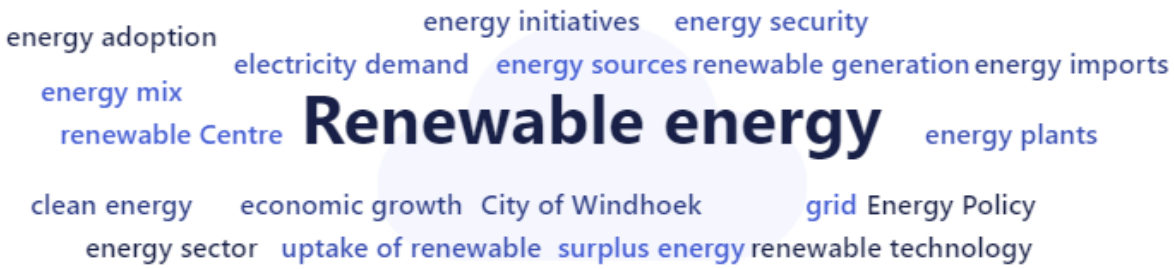
37 Responses

Significantly higher, causing financial strain	10
Slightly higher, manageable impact	13
Roughly equal, no significant impact	6
Slightly lower, leading to potential savings	5
Significantly lower, causing inefficiencies and higher per-unit costs	3



11. Is there any additional information or insights you would like to share regarding electricity demand, economic growth, and renewable energy integration in the City of Windhoek?

37 Responses



APPENDIX C: NOTES ON RAW DATA PRE-PROCESSING

This section explains/describes the data quality issues encountered on the CoW historical energy consumption raw data and the actions taken.

Cleaning Action Performed	Description
Renaming columns	Column names were standardised by removing problematic characters and spaces to ensure consistency and compatibility with various data processing libraries. For example 'E. Peak Rate (N$/kWh)' was renamed to E_Peak_Rate_N$/kWh. These changes improved the clarity and consistency of the column names, making the dataset easier to work with and reducing the risk of errors during data processing and analysis. The cleaned and standardised column names were then used in all subsequent steps of data preprocessing and modeling.
Combining Relevant Columns due to tariff unbundling	In 2020, the City of Windhoek undertook a tariff unbundling exercise, which resulted in the separation of electricity tariffs into distinct components to enhance transparency and better reflect the actual cost structure of electricity supply. This also marked the transition from a dual-season billing system, with high and low season rates, to a single-season billing approach. The tariffs were unbundled into two main categories: Energy and Losses **Energy Component:** This part of the tariff represents the cost associated with the generation and supply of electrical energy (UNESCAP, n.d). It includes costs related to the production of electricity, transmission, and distribution up to the point of consumption. The energy component was further divided based on the time of use, with different rates for peak, standard, and off-peak periods to encourage efficient energy usage and load management. **Losses Component:** This component accounts for the energy losses that occur during the transmission and distribution of electricity. These losses are an inevitable part of the electrical system and need to be factored into the tariff structure to ensure the financial sustainability of the utility. This component was also divided based on the time of use rates. Prior to 2020, the data values were bundled and did not separately account for the losses component; it was included in the total energy values. To ensure uniformity for modeling and to accurately reflect the post-2020 tariff unbundling into energy and losses components, these columns were added. This allowed for a consistent dataset structure, facilitating more accurate and effective modeling. The specific columns that were created and combined are as follows: **Energy Peak Rate (N$/kWh):** Combined values from 'E. Peak Rate (N$/kWh)' (representing energy component) and 'L. Peak Rate (N$/kWh)' (representing losses component).

	Energy Peak Total (N$): Combined values from 'E. Peak Total (N$)' and 'L. Peak Total (N$)'. **Energy Standard Rate (N$/kWh):** Combined values from 'E. Standard Rate (N$/kWh)' and 'L. Standard Rate (N$/kWh)'. **Energy Standard Total (N$):** Combined values from 'E. Standard Total (N$)' and 'L. Standard Total (N$)'. **Energy Offpeak Rate (N$/kWh):** Combined values from 'E. Offpeak Rate (N$/kWh)' and 'L. Offpeak Rate (N$/kWh)'. **Energy Offpeak Total (N$):** Combined values from 'E. Offpeak Total (N$)' and 'L. Offpeak Total (N$)'. These original columns, where values were added from, were then dropped to reduce redundancy and simplify the dataset, ensuring a more streamlined and efficient structure for analysis.
Missing values	These are observations in the dataset that have no recorded value. These values were handled by imputing (replacing missing values with estimated values based on other observations (Emmanuel, et al., 2021)). This was implemented using the **SimpleImputer** class from the scikit-learn library in python. Imputation was selected for its broad usage, effectiveness, simplicity, and straightforward implementation. **Imputation Strategies** **Numerical Columns:** Missing values in numerical columns were imputed using the mean value of each respective column. This method was chosen to maintain the statistical integrity and overall distribution of the data, ensuring that no significant biases were introduced during the imputation process. **Categorical Columns:** For categorical columns, missing values were imputed using the mode (most frequent category). This approach preserves the categorical distribution and ensures consistency in the dataset, essential for maintaining the quality of subsequent analyses.
Duplicate rows	These are observations that are identical to other observations in the dataset. The dataset did not have any duplicate rows.
Outliers	These are observations that significantly deviate from the rest of the dataset (Aggarwal, 2017). Scatter plots were used to visually identify observations that significantly deviated from the rest of the dataset. Upon detection, outliers were scrutinised further to understand their origins. Many outliers were traced back to the original data source (manual invoices) where corrections were made due to human errors or misspellings during data entry.
Dropping irrelevant columns	Several columns deemed not meaningful or redundant were dropped to streamline the dataset, making it more focused and relevant for the task of energy consumption forecasting. Hence, the following columns/features were dropped.

	Service Charges Units & S. Charge Rate (N$/Month): These columns relate to fixed monthly charges rather than variable consumption-based charges, and thus do not provide meaningful information for forecasting energy consumption. **ECB Levies & NEF Levies:** These levies are additional charges imposed by regulatory authorities and do not directly correlate with energy consumption patterns or trends, making them irrelevant for consumption forecasting. **Losses Peak in kWh, Losses Standard in kWh, Losses Off peak in kWh:** While these columns represent energy losses, they do not directly impact the actual consumption data required for forecasting purposes. The focus of the forecasting model is on the total consumption rather than losses, which are already accounted for in the combined total energy values.

APPENDIX D: IMPLEMENTATION CODE SNIPPETS

This appendix shows code snippets for selected key parts of implementation of the work presented in this dissertation. The complete implementation code is available on Gitlab and can be accessed using the following link: *https://gitlab.com/DianaNambahu/ndakalako-diana-nambahu-dissertation-full-python-code* (Note: If the link doesn't open when clicked, please copy and paste it directly into your web browser's to access the repository).

D.1: Combining columns due to tariff unbundling

```
# Add the values of columns "E. Standard Rate (N$/kWh)" and "L. Standard Rate (N$/kWh)"
# and store them in a new column "Energy_Standard_Rate_(N$/kWh)"
cow_energy_consumption['Energy_Standard_Rate_(N$/kWh)'] = cow_energy_consumption
['E._Standard_Rate_(N$/kWh)'] + cow_energy_consumption['L._Standard_Rate_(N$/kWh)']

# Display the first few rows to verify the changes
print(cow_energy_consumption[['E._Standard_Rate_(N$/kWh)', 'L._Standard_Rate_(N$/kWh)',
                              'Energy_Standard_Rate_(N$/kWh)']].head())
```

```
   E._Standard_Rate_(N$/kWh)  L._Standard_Rate_(N$/kWh)  \
0                     1.3275                     0.1495
1                     1.3275                     0.1495
2                     1.3275                     0.1495
3                     1.3275                     0.1495
4                     1.3275                     0.1495

   Energy_Standard_Rate_(N$/kWh)
0                          1.477
1                          1.477
2                          1.477
3                          1.477
4                          1.477
```

D.2: Handling missing values

```
# Check for missing values
print("\nMissing values:")
print(cow_energy_consumption.isnull().sum())
```

```
Missing values:
Billing_month                    0
Billing_Season                   0
Max._Demand_Date_&_Time          0
PF_@_Max.Dem                     3
S._Charge_Total_(N$)             0
Point_of_Supply__Charge          0
Max._Demand_in_kVA               1
Max._Demand_Rate_(N$/kVA)        0
Max._Demand_Total_(N$)           1
Network_Access_in_kVA            0
N._Access_Rate_(N$/kVA)          0
N._Access_Total_(N$)             0
Demand_Penalty_in_kVA            0
D._Penalty_Rate_(N$/kVA)         0
D._Penalty_Total_(N$)            0
Energy_Peak_in_kWh               1
Energy_Standard_in_kWh           1
Energy_Offpeak_in_kWh            1
Total_Energy_in_kWh              1
Total_Energy_Amount_(N$)         0
Reliability_Charge               0
```

```
# Define imputer for numerical columns (fill missing values with mean)
numerical_imputer = SimpleImputer(strategy='mean')
# Define numerical and categorical columns with missing values
numerical_columns_with_missing = ['Max._Demand_in_kVA', 'Max._Demand_Total_(N$)',
                                  'Energy_Peak_in_kWh','Energy_Standard_in_kWh',
                                  'Energy_Offpeak_in_kWh', 'Total_Energy_in_kWh',
                                  'VAT_@_15%_(N$)', 'ECB_Levies_Total_(N$)', 'NEF_Levies_Total_(N$)',
                                  'Total_Amount_for_the_month_(N$)', 'Energy_Peak_Total_(N$)',
                                  'Energy_Standard_Total_(N$)', 'PF_@_Max.Dem',
                                  'Energy Offpeak Total (N$)']
# Impute missing values for numerical columns
cow_energy_consumption[numerical_columns_with_missing] = numerical_imputer.fit_transform
(cow_energy_consumption[numerical_columns_with_missing])
# Check for any remaining missing values
print("\nMissing values after imputation:")
print(cow_energy_consumption[numerical_columns_with_missing].isnull().sum())
```

```
Missing values after imputation:
Max._Demand_in_kVA                0
Max._Demand_Total_(N$)            0
Energy_Peak_in_kWh                0
Energy_Standard_in_kWh            0
Energy_Offpeak_in_kWh             0
Total_Energy_in_kWh               0
VAT_@_15%_(N$)                    0
ECB_Levies_Total_(N$)             0
NEF_Levies_Total_(N$)             0
Total_Amount_for_the_month_(N$)   0
Energy_Peak_Total_(N$)            0
Energy_Standard_Total_(N$)        0
PF_@_Max.Dem                      0
Energy Offpeak Total (N$)         0
dtype: int64
```

D.3: Handling outliers

```
# Dictionary containing outlier values and their correct replacements
outlier_replacements = {
    (58, 'NEF_Levies_Total_(N$)'): 1217535.82,
    (73, 'NEF_Levies_Total_(N$)'): 1210788.62,
    (117, 'NEF_Levies_Total_(N$)'): 799692.76,
    (97, 'ECB_Levies_Total_(N$)'): 1291392.00,
    (100, 'ECB_Levies_Total_(N$)'): 1286196.00,
    }

# Update DataFrame with outlier replacements
for (row_index, column_name), correct_value in outlier_replacements.items():
    cow_energy_consumption.at[row_index, column_name] = correct_value
```

D.4: Data interpolation process

```
import pandas as pd

# Convert date columns to datetime format
e_consumption['Date'] = pd.to_datetime(e_consumption['Date'], format='%d/%m/%Y')
gdp_pop['Date'] = pd.to_datetime(gdp_pop['Date'], format='%Y/%m/%d')
re_energy['Financial year'] = pd.to_datetime(re_energy['Financial year'], format='%d/%m/%Y')

# Create a complete date range from the min to max date in e_consumption
full_date_range = pd.date_range(start=e_consumption['Date'].min(), end=e_consumption['Date'].max(),
                                freq='MS')

# Reindex and interpolate gdp_pop data
gdp_pop = gdp_pop.set_index('Date').reindex(full_date_range).interpolate(method='linear')
gdp_pop.index.name = 'Date'
gdp_pop = gdp_pop.reset_index()

# Reindex and interpolate re_energy data
re_energy = re_energy.set_index('Financial year').reindex(full_date_range).interpolate(method='linear')
re_energy.index.name = 'Date'
re_energy = re_energy.reset_index()

# Merge datasets on Date column
merged_df = pd.merge(e_consumption, gdp_pop, on='Date', how='left')
merged_df = pd.merge(merged_df, re_energy, on='Date', how='left')
```

D.5: Data loading and preprocessing for modelling

```
# Step 1: Load and preprocess the data
df = pd.read_csv('merged_df_cleaned.csv')

# Convert the 'Date' column to datetime
df['Date'] = pd.to_datetime(df['Date'], format='%d/%m/%Y')

# Set the 'Date' column as the index
df.set_index('Date', inplace=True)

# Normalize all relevant columns
scaler = MinMaxScaler(feature_range=(0, 1))
scaled_data = scaler.fit_transform(df)

# Convert the scaled data back to a DataFrame
scaled_df = pd.DataFrame(scaled_data, columns=df.columns, index=df.index)
```

```
# Function to create sequences
def create_sequences(data, seq_length):
    X, y = [], []
    for i in range(len(data) - seq_length):
        X.append(data[i:i + seq_length].flatten())
        y.append(data[i + seq_length, 0])  # Predict the 'Consumption' column
    return np.array(X), np.array(y)

seq_length = 12  # Use past 12 months for prediction
data = scaled_df.values
X, y = create_sequences(data, seq_length)

# Split the data into training and testing sets
split_ratio = 0.85
train_size = int(len(X) * split_ratio)
X_train, X_test = X[:train_size], X[train_size:]
y_train, y_test = y[:train_size], y[train_size:]
```

D.6: ML Models Training

```
# LSTM Model
def create_sequences_lstm(data, seq_length):
    X, y = [], []
    for i in range(len(data) - seq_length):
        X.append(data[i:i + seq_length])
        y.append(data[i + seq_length, 0])
    return np.array(X), np.array(y)

X_lstm, y_lstm = create_sequences_lstm(scaled_df.values, seq_length)

# Split the data into training and testing sets
X_train_lstm, X_test_lstm = X_lstm[:train_size], X_lstm[train_size:]
y_train_lstm, y_test_lstm = y_lstm[:train_size], y_lstm[train_size:]

# Build and train the LSTM model
model = Sequential()
model.add(LSTM(units=50, return_sequences=True, input_shape=(seq_length, X_train_lstm.shape[2])))
model.add(LSTM(units=50))
model.add(Dense(units=1))

model.compile(optimizer='adam', loss='mean_squared_error')
model.summary()

history = model.fit(X_train_lstm, y_train_lstm, epochs=50, batch_size=1, validation_split=0.2)

y_pred_lstm = model.predict(X_test_lstm)
y_test_inv_lstm = scaler.inverse_transform(np.concatenate((y_test_lstm.reshape(-1, 1), X_test_lstm[:, -1, 1:]), axis=1))[:, 0]
y_pred_inv_lstm = scaler.inverse_transform(np.concatenate((y_pred_lstm, X_test_lstm[:, -1, 1:]), axis=1))[:, 0]

# Plot the training and validation loss
plt.plot(history.history['loss'], label='Training loss')
plt.plot(history.history['val_loss'], label='Validation loss')
plt.legend()
plt.show()

# Evaluate and plot LSTM
results.append(evaluate_and_plot("LSTM Model", y_test_inv_lstm, y_pred_inv_lstm))
```

```
# Random Forest Regressor
param_grid_rf = {
    'n_estimators': [100, 200, 300],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}

tscv_rf = TimeSeriesSplit(n_splits=5)
rf_model = RandomForestRegressor(random_state=42)

grid_search_rf = GridSearchCV(estimator=rf_model, param_grid=param_grid_rf, cv=tscv_rf, scoring='neg_mean_squared_error',
grid_search_rf.fit(X_train, y_train)

best_model_rf = grid_search_rf.best_estimator_

y_pred_rf = best_model_rf.predict(X_test)

X_test_last_rf = X_test[:, -len(df.columns) + 1:]

y_test_inv_rf = scaler.inverse_transform(np.concatenate((y_test.reshape(-1, 1), X_test_last_rf), axis=1))[:, 0]

y_pred_inv_rf = scaler.inverse_transform(np.concatenate((y_pred_rf.reshape(-1, 1), X_test_last_rf), axis=1))[:, 0]

results.append(evaluate_and_plot("Random Forest Regressor", y_test_inv_rf, y_pred_inv_rf))
```

```python
# Gradient Boosting Regressor

# Hyperparameter tuning using GridSearchCV
param_grid_gb = {
    'n_estimators': [50, 100, 200, 300],
    'learning_rate': [0.001, 0.01, 0.05, 0.1, 0.2],
    'max_depth': [3, 4, 5, 6],
    'min_samples_split': [2, 5, 10, 20],
    'min_samples_leaf': [1, 2, 4, 10],
    'subsample': [0.7, 0.8, 0.9, 1.0]
}

tscv_gb = TimeSeriesSplit(n_splits=5)
gb_model = GradientBoostingRegressor(random_state=42)

grid_search_gb = GridSearchCV(estimator=gb_model, param_grid=param_grid_gb, cv=tscv_gb, scoring='neg_mean_squared_error',
grid_search_gb.fit(X_train, y_train)

# Best model from GridSearch
best_model_gb = grid_search_gb.best_estimator_

# Evaluate the best model
y_pred_gb = best_model_gb.predict(X_test)

# Prepare data for inverse transformation
X_test_last_gb = X_test[:, -len(df.columns) + 1:]  # The last time step's features, except for the target

# Inverse transform y_test
y_test_inv_gb = scaler.inverse_transform(np.concatenate((y_test.reshape(-1, 1), X_test_last_gb), axis=1))[:, 0]

# Inverse transform y_pred
y_pred_inv_gb = scaler.inverse_transform(np.concatenate((y_pred_gb.reshape(-1, 1), X_test_last_gb), axis=1))[:, 0]
```

```python
# LightGBM Regressor
lgb_model = lgb.LGBMRegressor(n_estimators=100, learning_rate=0.05, random_state=42)
lgb_model.fit(X_train, y_train)

y_pred_lgb = lgb_model.predict(X_test)

X_test_last_lgb = X_test[:, -len(df.columns) + 1:]

y_test_inv_lgb = scaler.inverse_transform(np.concatenate((y_test.reshape(-1, 1), X_test_last_lgb), axis=1))[:, 0]

y_pred_inv_lgb = scaler.inverse_transform(np.concatenate((y_pred_lgb.reshape(-1, 1), X_test_last_lgb), axis=1))[:, 0]

results.append(evaluate_and_plot("LightGBM Regressor", y_test_inv_lgb, y_pred_inv_lgb))
```

D.7: Statistical Models Training

```
# ARIMA Model
# Fit ARIMA(3, 1, 3) model on the training set
model_arima = sm.tsa.ARIMA(train['Consumption'], order=(3, 1, 3))
results_arima = model_arima.fit()

# Generate predictions for the test set
start = train_size
end = len(df) - 1
predictions_arima = results_arima.predict(start=start, end=end, typ='levels')

# Actual values from the test set
actual_values_arima = test['Consumption']

# Calculate metrics
mae_arima = mean_absolute_error(actual_values_arima, predictions_arima)
mse_arima = mean_squared_error(actual_values_arima, predictions_arima)
rmse_arima = np.sqrt(mse_arima)
r2_arima = r2_score(actual_values_arima, predictions_arima)
mape_arima = np.mean(np.abs((actual_values_arima - predictions_arima) / actual_values_arima)) * 100
```

```
# SARIMA Model
# Use auto_arima to find the best model parameters
s = 12  # seasonal period
auto_model_sarima = pm.auto_arima(train['Consumption'], seasonal=True, m=s, trace=True, error_action='ignore',

# Fit the best SARIMA model on the training data
sarima_model = auto_model_sarima.fit(train['Consumption'])

# Forecasting
forecast_steps = len(test)
forecast_sarima = sarima_model.predict(n_periods=forecast_steps)
forecast_index_sarima = test.index

# Calculate metrics
mae_sarima = mean_absolute_error(test['Consumption'], forecast_sarima)
mse_sarima = mean_squared_error(test['Consumption'], forecast_sarima)
rmse_sarima = np.sqrt(mse_sarima)
r2_sarima = r2_score(test['Consumption'], forecast_sarima)
mape_sarima = np.mean(np.abs((test['Consumption'] - forecast_sarima) / test['Consumption'])) * 100
```

D.8: Model evaluation

```python
# Function to evaluate and plot results
def evaluate_and_plot(model_name, y_test_inv, y_pred_inv):
    mse = mean_squared_error(y_test_inv, y_pred_inv)
    mae = mean_absolute_error(y_test_inv, y_pred_inv)
    rmse = np.sqrt(mse)
    r2 = r2_score(y_test_inv, y_pred_inv)
    mape = np.mean(np.abs((y_test_inv - y_pred_inv) / y_test_inv)) * 100

    print(f"{model_name} Evaluation:")
    print(f"Mean Absolute Error (MAE): {mae}")
    print(f"Mean Squared Error (MSE): {mse}")
    print(f"Root Mean Squared Error (RMSE): {rmse}")
    print(f"R^2 Score: {r2}")
    print(f"R^2 Score (%): {r2 * 100}")
    print(f"Mean Absolute Percentage Error (MAPE): {mape}\n")

    plt.figure(figsize=(14, 5))
    plt.plot(y_test_inv, label='Actual')
    plt.plot(y_pred_inv, label='Predicted')
    plt.title(f'{model_name} - Actual vs Predicted')
    plt.legend()
    plt.show()

    return {'Model': model_name, 'MAE': mae, 'MSE': mse, 'RMSE': rmse, 'R2': r2, 'MAPE': mape}
```

D.9: Model hybridisation

```python
# Create a new dataset for the meta-learner
stacked_predictions = np.column_stack((y_pred_inv_rf[:len(test_arima)], arima_predictions))

# Define actual values
actual_values = test_arima['Consumption']

# Train the meta-learner (Linear Regression)
meta_learner = LinearRegression()
meta_learner.fit(stacked_predictions, actual_values)

# Final predictions
final_predictions = meta_learner.predict(stacked_predictions)

# Evaluate the hybrid model
hybrid_mae = mean_absolute_error(actual_values, final_predictions)
hybrid_mse = mean_squared_error(actual_values, final_predictions)
hybrid_rmse = np.sqrt(hybrid_mse)
hybrid_mape = np.mean(np.abs((actual_values - final_predictions) / actual_values)) * 100
```

```
# Create a new dataset for the meta-learner
stacked_predictions = np.column_stack((y_pred_inv_rf[:len(test_arima)], arima_predictions))

# Define actual values
actual_values = test_arima['Consumption']

# Train the meta-learner (Gradient Boosting Regressor)
meta_learner = GradientBoostingRegressor(n_estimators=100, random_state=2)
meta_learner.fit(stacked_predictions, actual_values)

# Final predictions
final_predictions = meta_learner.predict(stacked_predictions)

# Evaluate the hybrid model
hybrid_mae = mean_absolute_error(actual_values, final_predictions)
hybrid_mse = mean_squared_error(actual_values, final_predictions)
hybrid_rmse = np.sqrt(hybrid_mse)
hybrid_mape = np.mean(np.abs((actual_values - final_predictions) / actual_values)) * 100
```

D.10: Electricity consumption forecast

```
# Train the meta-learner (Gradient Boosting Regressor)
meta_learner = GradientBoostingRegressor(n_estimators=100, random_state=2)
meta_learner.fit(stacked_predictions, actual_values)

# Helper function to create future sequences
def create_future_sequences(data, seq_length, num_steps):
    future_sequences = []
    current_seq = data[-seq_length:]
    for _ in range(num_steps):
        future_sequences.append(current_seq.flatten())
        next_pred = best_model_rf.predict(current_seq.flatten().reshape(1, -1))
        next_seq = np.roll(current_seq, -1, axis=0)
        next_seq[-1] = np.concatenate([next_pred, current_seq[-1, 1:]])
        current_seq = next_seq
    return np.array(future_sequences)

# Predict the next two years (24 months)
future_steps = 24
```

D.11: Financial losses calculations for Maximum Demand Discrepancies

```
# Calculate the total financial losses
total_financial_losses = df2['Loss_due_to_Higher_Demand'].sum()

print(f'Total Financial Losses: {total_financial_losses} N$')
```

```
Total Financial Losses: 288242176.65291804 N$
```

```
# Calculate the average monthly financial loss
average_monthly_loss = total_financial_losses / len(df2['Billing_month'].unique())

print(f'Average Monthly Financial Loss: {average_monthly_loss:.2f} N$')
```

```
Average Monthly Financial Loss: 2343432.33 N$
```

D.12: Overall Financial Implications of Maximum Demand Discrepancies

```
# Total financial implications
total_max_demand_charge = df2['Max._Demand_Total_(N$)'].sum()
total_notified_demand_charge = df2['N._Access_Total_(N$)'].sum()
total_penalty_charge = df2['D._Penalty_Total_(N$)'].sum()
print("\nTotal Financial Implications:")
print(f"Total Max Demand Charge (N$): {total_max_demand_charge:,.2f}")
print(f"Total Notified Demand Charge (N$): {total_notified_demand_charge:,.2f}")
print(f"Total Penalty Charge (N$): {total_penalty_charge:,.2f}")
# Monthly averages
monthly_avg_max_demand_charge = df2.groupby(merged_df['Date'].dt.to_period('M'))['Max._Demand_Total_(N$
monthly_avg_notified_demand_charge = df2.groupby(merged_df['Date'].dt.to_period('M'))['N._Access_Total_
monthly_avg_penalty_charge = df2.groupby(merged_df['Date'].dt.to_period('M'))['D._Penalty_Total_(N$)'].
print("\nMonthly Averages:")
print(f"Average Monthly Max Demand Charge (N$): {monthly_avg_max_demand_charge:,.2f}")
print(f"Average Monthly Notified Demand Charge (N$): {monthly_avg_notified_demand_charge:,.2f}")
print(f"Average Monthly Penalty Charge (N$): {monthly_avg_penalty_charge:,.2f}")
```

```
Total Financial Implications:
Total Max Demand Charge (N$): 1,614,439,195.64
Total Notified Demand Charge (N$): 1,792,427,216.42
Total Penalty Charge (N$): 1,115,976.15

Monthly Averages:
Average Monthly Max Demand Charge (N$): 13,125,521.92
Average Monthly Notified Demand Charge (N$): 14,572,579.00
Average Monthly Penalty Charge (N$): 9,072.98
```

APPENDIX E: A REFLECTION ON MY RESEARCH JOURNEY

The dual nature of the student experience, often described as "here yet not here," resonates deeply with my own journey through this research. It embodies the paradox of feeling both confident and uncertain at various points—moments when I felt in control and on top of my game, and others when I doubted my entire direction. This duality powerfully mirrors my journey of fulfilling this dissertation, where the highs and lows of learning were ever-present.

Embarking on this research journey has been both challenging and deeply fulfilling. The process has not only strengthened my technical expertise but also shaped my perspective on energy sustainability and urban planning. Through the different stages of this study, I have encountered moments of inspiration, learning, and professional growth.

The inspiration for my research began when I observed the inefficiencies in energy management in rapidly growing urban areas like Windhoek. With population growth, economic development, and renewable energy integration reshaping urban power dynamics, I realised the pressing need for accurate electricity demand forecasting. My curiosity was sparked, and I knew this was an area worth exploring, particularly for the City of Windhoek.

The process of collecting data for this research was perhaps the most difficult phase. The historical electricity consumption data I needed was scattered across multiple sources, both electronic and hard copies, dating back over a decade. Simultaneously, I collected data on economic indicators such as GDP and population growth, recognizing that these factors play a crucial role in electricity demand.

The process of data cleaning and model implementation was equally demanding. Through rigorous experimentation, I identified **Random Forest** and **ARIMA** as the best-performing individual models. The integration of these models, using a stacking approach with **Gradient Boosting Regressor** as the meta-learner, formed the core of my hybrid model. This stage of the research was particularly engaging, as it allowed me to immerse myself in model training, hyper parameter tuning, and the intricate balance of machine learning with traditional statistical techniques. It was fascinating to see how these methods, when combined, could offer superior predictions.

I am immensely grateful to my supervisor, Dr. Richard Maliwatu, for his unwavering support, as well as the City of Windhoek for facilitating my research. I also deeply appreciate my friends and classmates—Angula, Julia, Teopolina, and Dorian—for making this journey both memorable and manageable. They turned sleepless nights into shared fun experiences, which helped me push through the toughest phases of the research.

As I look ahead, I am excited about the potential impact of my findings on electricity demand management in Windhoek and beyond. This journey has equipped me with invaluable skills and insights that I hope will contribute to real-world solutions in the future.