

# **Building Household Electricity Load Profile Using Machine Learning**



By

**Arisha Saeed Akkas**

2018-NUST-MSCS-8 00000275764

Supervisor

**Dr. Fahad Javed**


A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science  
in Computer Science (MS CS)

Department of Computing  
School of Electrical Engineering and Computer Science (SEECS)  
National University of Sciences and Technology (NUST)  
Islamabad, Pakistan.

(July 2022)

## **THESIS ACCEPTANCE CERTIFICATE**

Certified that final copy of MS/MPhil thesis entitled "Building household electricity load profiles using machine learning" written by ARISHA AKKAS, (Registration No 00000275764), of SEecs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature:  \_\_\_\_\_

Name of Advisor: Dr. Fahad Javed

Date: 26-Mar-2021

HOD/Associate Dean: \_\_\_\_\_

Date: \_\_\_\_\_

Signature (Dean/Principle): \_\_\_\_\_

Date: \_\_\_\_\_

## Approval

It is certified that the contents and form of the thesis entitled “Building Household Electricity Load Profiles Using Machine Learning” submitted by Arisha Akkas, have been found satisfactory for the requirement of the degree.

Advisor: Dr. Fahad Javed

Signature: \_\_\_\_\_

Date: 26-Mar-2021

Committee Member 1: Dr. Safdar Abbas Khan

Signature: \_\_\_\_\_

Date: 26-Mar-2021

Committee Member 2: Dr. Imran Mahmood Hashmi

Signature: \_\_\_\_\_

Date: 4-May-2021

Committee Member 3: Dr. Syed Taha Ali

Signature: \_\_\_\_\_

Date: 25-Mar-2021

## **Dedication**

I dedicate this thesis to the people who have been deprived of clean energy, because of the negligence of the public. Also, to my pets and other animals, for whom we have made this earth a living hell, thanks to the CO<sub>2</sub> emissions.

## Certificate of Originality

I hereby declare that this submission titled "Building Household Electricity Load Profiles Using Machine Learning" is my own work. To the best of my knowledge, it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation, and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Author's Name: Arisha Saeed Akkas

Signature: 

## **Acknowledgement**

I was only able to make it this far because my parents didn't make me think twice about my decisions. To them, I not only owe this thesis, but my life. There are so many people to whom I owe a thank you, because one way or another they contributed towards my research. Starting with my supervisor, who made it all happen. His brilliant idea and thorough guidance compelled me towards this research. He has the greatest endurance, that made him cope with all my setbacks and yet, push me to enough so I make it to the finish line. And finally, my husband, who was the very first person to encourage me to code, to put logic into syntax. I dedicate this effort to you guys.

# Table of Contents

<b>List of Abbreviations and Symbols .....</b>	<b>ix</b>
<b>List of Tables .....</b>	<b>x</b>
<b>List of Figures.....</b>	<b>xii</b>
<b>Abstract.....</b>	<b>xiv</b>
<b>1. Introduction.....</b>	<b>1</b>
1.1. Introduction .....	1
1.2. Electricity Consumption Data with Smart Meters .....	2
1.2.1. Smart Meters in Pakistan .....	2
1.3. Building Energy Models (BEMs) .....	2
1.3.1. Approaches for Developing BEMs.....	3
1.4. Research Objectives .....	4
1.5. Overview of the Proposed Approach .....	4
1.6. Structure of Thesis .....	5
<b>2. Literature Review .....</b>	<b>6</b>
2.1. Chapter Overview .....	6
2.2. Background .....	6
2.2.1. Electric Load Profiles Based on Historic Data .....	6
2.2.2. Electricity Load Profiles Based on Socio-Economic Parameters .....	11
2.3. Conclusion and Comparative Summary.....	15
<b>3. Methodology .....</b>	<b>17</b>
3.1. Datasets .....	18
3.1.1. REDD.....	19
3.1.2. PRECON.....	23
3.1.3. REFIT .....	25

3.1.4.	Sky Electric Dataset .....	27
3.1.5.	Dataset Summary .....	28
3.2.	Pre-Processing .....	29
3.2.1.	Finalizing Data Features .....	29
3.2.2.	Finalizing Data Granularity .....	32
3.2.3.	Altering Data Representation (Converting to Binary Representation) .....	34
3.3.	Post Analysis .....	35
3.3.1.	Time Window Segmentation .....	35
3.4.	Machine Learning and Deep Learning .....	37
3.4.1.	Random Forest Classifier .....	38
3.4.2.	Hidden Markov Model .....	38
3.4.3.	Artificial Neural Nets (ANN) .....	39
3.5.	Evaluation Metrics .....	39
<b>4.</b>	<b>Results .....</b>	<b>41</b>
4.1.	Results Compilation for REFIT Dataset .....	41
4.1.1.	Without Socio Economic (SE) Features .....	42
4.1.1.1.	Washing Machine .....	42
4.1.1.2.	Dishwasher .....	43
4.1.1.3.	Television .....	45
4.1.1.4.	Microwave .....	46
4.1.1.5.	Kettle .....	48
4.1.2.	With Socio Economic (SE) Features .....	49
4.1.2.1.	Washing Machine .....	49
4.1.2.2.	Dishwasher .....	51
4.1.2.3.	Television .....	52



4.1.2.4.	Microwave.....	54
4.1.2.5.	Kettle .....	55
4.2.	Results Compilation for Sky Electric Dataset.....	57
4.2.1.	Microwave .....	57
4.2.2.	Power Socket .....	59
4.2.3.	High Voltage Bulb .....	61
4.2.4.	AC (Drawing Room).....	63
4.2.5.	AC (Bedroom) .....	65
4.3.	Comparison of Approaches .....	66
<b>5.</b>	<b>Discussions .....</b>	<b>69</b>
5.1.	Summary of the Findings .....	69
5.2.	Comparison with Existing Work.....	70
5.3.	Limitations .....	70
<b>6.</b>	<b>Conclusions.....</b>	<b>72</b>
6.1.	Summary of Research Contributions .....	72
6.2.	Future Work .....	73
<b>7.</b>	<b>References.....</b>	<b>74</b>

## List of Abbreviations and Symbols

### Abbreviations

BEMs	Building Energy Model
SE	Socio Economic
RF	Random Forest
HMM	Hidden Markov Model
ANN	Artificial Neural Network
REDD	Reference Energy Disaggregation Dataset
REFIT	Electrical Load Measurement Data
PRECON	Pakistan's Residential Electricity Consumption

## List of Tables

Table 2.1: Side by side comparison of different electricity load profile generation strategies in the literature .....	15
Table 3.1: Summarized Comparison of the Datasets .....	28
Table 3.2: Selected Features from REFIT and Sky Electric .....	29
Table 4.1: Summary of Details Covered Under Results .....	41
Table 4.2: Random Forest Classifier for Predicting Washing Machine State – No SE Features .	42
Table 4.3: Hidden Markov Model for Predicting Washing Machine State – No SE Features .....	43
Table 4.4: ANN for Predicting Washing Machine State - No SE Features .....	43
Table 4.5: Random Forest Classifier for Predicting Dishwasher State – No SE Features .....	44
Table 4.6: Hidden Markov Model for Predicting Dishwasher State – No SE Features .....	44
Table 4.7: ANN for Predicting Dishwasher State - No SE Features .....	45
Table 4.8: Random Forest Classifier for Predicting Television State – No SE Features .....	45
Table 4.9: Hidden Markov Model for Predicting Television State – No SE Features .....	46
Table 4.10: ANN for Predicting Television State - No SE Features .....	46
Table 4.11: Random Forest Classifier for Predicting Microwave State – No SE Features .....	47
Table 4.12: Hidden Markov Model for Predicting Microwave State – No SE Features .....	47
Table 4.13: ANN for Predicting Microwave State - No SE Features .....	48
Table 4.14: Random Forest Classifier for Predicting Kettle State – No SE Features .....	48
Table 4.15: Hidden Markov Model for Predicting Kettle State – No SE Features .....	49
Table 4.16: ANN for Predicting Kettle State - No SE Features .....	49
Table 4.17: Random Forest Classifier for Predicting Washing Machine State – SE Features .....	50
Table 4.18: Hidden Markov Model for Predicting Washing Machine State – SE Features .....	50
Table 4.19: ANN for Predicting Washing Machine State - SE Features .....	51
Table 4.20: Random Forest Classifier for Predicting Dishwasher State – SE Features .....	51
Table 4.21: Hidden Markov Model for Predicting Dishwasher State – SE Features .....	52
Table 4.22: ANN for Predicting Dishwasher State - SE Features .....	52
Table 4.23: Random Forest Classifier for Predicting Television State – SE Features .....	53
Table 4.24: Hidden Markov Model for Predicting Television State – SE Features .....	53
Table 4.25: ANN for Predicting Television State - SE Features .....	54
Table 4.26: Random Forest Classifier for Predicting Microwave State – SE Features .....	54

Table 4.27: Hidden Markov Model for Predicting Microwave State – No SE Features .....	55
Table 4.28: ANN for Predicting Microwave State - SE Features.....	55
Table 4.29: Random Forest Classifier for Predicting Kettle State – SE Features .....	56
Table 4.30: Hidden Markov Model for Predicting Kettle State – SE Features .....	56
Table 4.31: ANN for Predicting Kettle State - SE Features .....	57
Table 4.32: Random Forest Classifier for Predicting Microwave State .....	58
Table 4.33: Hidden Markov Model for Predicting Microwave State .....	58
Table 4.34: ANN for Predicting Microwave State .....	59
Table 4.35: Random Forest Classifier for Predicting Power Socket State .....	60
Table 4.36: Hidden Markov Model for Predicting Power Socket State .....	60
Table 4.37: ANN for Predicting Power Socket State .....	61
Table 4.38: Random Forest Classifier for Predicting HV Bulb State.....	62
Table 4.39: Hidden Markov Model for Predicting HV Bulb State.....	62
Table 4.40: ANN for Predicting HV Bulb State.....	63
Table 4.41: Random Forest Classifier for Predicting AC (Drawing Room) State .....	64
Table 4.42: Hidden Markov Model for Predicting AC (Drawing Room) State .....	64
Table 4.43: ANN for Predicting AC (Drawing Room) State.....	65
Table 4.44: Random Forest Classifier for Predicting AC (Bedroom) State .....	65
Table 4.45: Hidden Markov Model for Predicting AC (Bedroom) State .....	66
Table 4.46: ANN for Predicting AC (Bedroom) State .....	66
Table 4.47: Comparison of all REFIT Devices based on the Results from all Three Approaches and Type of the Dataset (with SE or Without SE). Results Shown in % .....	67
Table 4.48: A Comparison of all the Non-SE data (REFIT and Sky Electric) to compare the three approaches. Results Shown in % .....	68

## List of Figures

Figure 3.1: Pipeline to Generate Household electricity Load Profiles .....	18
Figure 3.2: REDD House 1; Refrigerator Consumption Pattern with Respect to Oven.....	20
Figure 3.3: REDD House 1; Refrigerator Consumption Pattern with Respect to Dishwasher.....	20
Figure 3.4: REDD House 1; Refrigerator Consumption Pattern with Respect to Washer Dryer .	20
Figure 3.5: REDD House 1; Lighting Device Consumption Pattern.....	21
Figure 3.6: REDD House 1; Device-Wise Consumption Analysis .....	21
Figure 3.7: REDD House 2; Device-Wise Consumption Analysis .....	22
Figure 3.8: REDD House 3; Device-Wise Consumption Analysis .....	22
Figure 3.9: REDD House 4; Device-Wise Consumption Analysis .....	22
Figure 3.10: Data Collection Architecture for PRECON (Nadeem & Arshad, 2019).....	23
Figure 3.11: Total Consumption Pattern of Refrigerator .....	25
Figure 3.12: Total Consumption Pattern of the Individual Monitored Devices PRECON .....	25
Figure 3.13: Schema of REFIT Data Collection (Murray et al., 2017) .....	26
Figure 3.14: Total Consumption Pattern of the Individual Monitored Devices REFIT .....	27
Figure 3.15: Total Consumption Pattern of the Individual Monitored Devices Sky Electric.....	28
Figure 3.16: Impact of the weekday on the electricity consumption of the devices - Sky Electric .....	30
Figure 3.17: Impact of the weekday on the electricity consumption of the devices – REFIT.....	30
Figure 3.18: Correlation Amongst All the Finalized Features – REFIT.....	31
Figure 3.19: Granularity of 1 minute - Sky Electric .....	32
Figure 3.20: Granularity of 5 minutes - Sky Electric.....	32
Figure 3.21: Granularity of 10 minutes - Sky Electric.....	33
Figure 3.22: Granularity of 15 minutes - Sky Electric.....	33
Figure 3.23: Granularity of 15 minutes - Sky Electric.....	33
Figure 3.24: Total Device-wise Consumption (Binarized) - Sky Electric.....	34
Figure 3.25: Total Device-wise Consumption (Binarized) – REFIT.....	35
Figure 3.26: REFIT Segmentation Window (12 am - 7 pm) .....	35
Figure 3.27: REFIT Segmentation Window (7 am - 10 pm).....	36
Figure 3.28: REFIT Segmentation Window (10 am - 3 pm) .....	36
Figure 3.29: REFIT Segmentation Window (3 pm - 8 pm).....	37

Figure 3.30: REFIT Segmentation Window (8 pm - 12 am) .....	37
Figure 5.1: Fall in the Precision of Kettle Due to Class Imbalance.....	71

## **Abstract**

The objective is to generate household electricity load-profiles by predicting device-wise electricity usage patterns using the time-series data, to trace the disutility of electricity. Further, the emphasis lies on how the socio-economic parameters of a household can combine with the time-series data to aid in better prediction. This is a comparative study representing a bottom-up model, with input granularity set to 10-min cycle power of 5 everyday household devices along with their associated timestamps as the building blocks and predicts, whether a device would be switched on or off at a given point in time.

The model is trained and validated on REFIT dataset, comprising of 20 houses along with the socio-economic features of each house. For comparison purpose, two datasets are created, with and without the socio-economic parameters. Results point towards the impact of socio-economic features and how they improved the prediction accuracy by a fine margin for each device, leading towards promising high-resolution electricity load profiles. Using the socio-economic features, we were able to predict the state of a device up to an accuracy of 97%, whereas without these features the accuracy was 76%.

**Keywords:** Electricity Load Profiles, Building Energy Models, Socio-economic Features, Granularity, Residential Data

# Chapter 1

## Introduction

### 1.1. Introduction

As the urbanization tends to drastically grow, we witness an exponential increase in the utilization of electricity. There are abundant literature references that point towards a positive correlation between growth and electricity consumption [1]. Electricity is the building block of development and economic progress within a country. It is the major driving factor for goods manufacturing and enabling capital and labor.

Shortage of electricity is one of the key effectors in bringing down a country's economy, hence causing precipitated growth [2][3][4][5] [6][7]. Studies like [8][9][10][11] have attributed shortage in electricity supply in developing countries to either poor infrastructures or low-income level. However, to a large extent, it is believed that environmental and energy policy design depends on the understanding this intermittent link [12]. Another issue that arises here to meet the rising demand, is the increased generation of electricity by means of coal and other fossil fuels which has led to a huge emission of carbon in the atmosphere leaving behind a horrendous impact [13]. Controlling the situation before it goes out of hand is the utmost need of the hour by taking one step at a time.

To understand what a perpetual catalyst in this problem could be, our focus lies in determining how residential sector is contributing towards electricity wastage leading to its shortage. Since it is one of the key sectors when it comes to consumption [14].

According to the literature, around one third of the total worldwide electricity is being consumed by the residential buildings, in figures this consumption goes from 27% to as high as 43% sometimes [15] and this demand keeps growing. The exponential growth that we witness in this trend is credited to the increased households because of the blooming rise in urbanization and because of the increase in the household appliances which attribute to an increased human comfort. These high living standards come at a cost of increased consumption and causing a prominent peak in the residential electricity consumption [13].



## **1.2. Electricity Consumption Data with Smart Meters**

With the introduction of smart meters gradually replacing the traditional ones, the key advantage that we have earned is that it can transfer the consumption information back to a data processing system [16]. This provides us with sufficient data that can be used to analyze the patterns of overall household consumption and in a long run disintegrate the total consumption and see how individual applications are being used. All these analyses can lead us towards improving the power utilization mechanism. And moving a step further from analyses this data can be leveraged and combined with multiple signal processing techniques from domains like stochastic analysis and artificial intelligence to further enhance the building energy models (BEMs). These reasons have made the smart meter technology ever so popular and their installation around the world has gotten increasingly high. Based on rough stats from 2016 around 70 million smart meters were installed across the USA and around 96 million across China [17].

### **1.2.1. Smart Meters in Pakistan**

Pakistan despite being an under-developed country with many issues in hand, has started the installation of smart meters as well, to avoid the electricity demand issues by controlling the overall shortage [18]. It is the need of the hour since in the recent years, the country has been on the edge with a shortage of 6000 MW electricity, causing a load-shedding of 10-14 hours in both the urban and the rural areas. Based on relatively recent statistics the demand of electricity in the country is around 25000 MW, however, the generation is only 18,900 MW. [19]. The mere installation of smart meters is not enough, a compilation of the data that can be further used to make some valuable prediction is required [20]. These kinds of datasets are hardly available for the under-developed countries, especially for the ones falling under the South Asian region. The ones that are available are small and provide insufficient data to analyze the situation and build a worthy solution to model the consumption behaviors [21]. However, in 2019 an effort was made to create PRECON (Pakistan Residential Electricity Consumption) dataset, it is considered large enough to provide details and perform research in the residential electricity consumption domain. [22].

## **1.3. Building Energy Models (BEMs)**

As the data sources of real-time electricity increase, the opportunity to utilize them and improve the current systems becomes a necessity. The key focus is to find the consumption patterns with high prediction accuracy, such that these load profiles can be used to enhance the current BEMs

and help in investigating the energy-saving potential of the buildings. Majorly there are two kinds of building electricity consumption behaviors, the ‘basic’ and the ‘variable’ behavior. Where basic is the behavior that depends on the area of the building while, variable behavior is dependent upon the occupancy [15]. The basic behavior standalone is insufficient to provide worthy insights on the electricity consumption patterns because, the pattern hardly relies on the area. As the person per building tends to grow, with each occupant carrying an independent behavior through which they consume the electricity, accounting the variable behavior when predicting the consumption patterns is essential.

Mastering the behavior of the occupants is the key to generate load profiles that can make significant contribution in improving the BEMs and help us in understanding the performance gap that exists in building design and operation. The energy consumption magnitude of a building is highly influenced by its design and the available technologies which in turn are associated with the social standards and behavioral characteristics of its occupants. As per research, single houses similar in size, location and envelope could have an energy consumption difference as large as 300% which is majorly because of the occupant behavior. As technology advances people focus solely on their comforts forgetting the deadly impact it holds on to the environment. Evidence proves that occupants residing in fully automated building tend to change the controls by manually altering them to attain the level of comfort they are aiming for [23]. To fully capture the occupant behavior, the key parameters to be recorded as input are the number of occupants residing, the appliances installed, the plug load schedules [24]. It is to be noted however, that detailed investigation of the occupants over a long period of time is not considered appropriate as it can harm the occupant’s privacy [15].

### **1.3.1. Approaches for Developing BEMs**

After covering the problem and its possible solution, comes the decision on how the model for generating and predicting the electric load profile should be developed. One of the major challenges is the non-linearity that exists amongst the load profiles; this can be triggered by a lot of factors including the occupancy behavior, the insulation of the building, the outdoor environmental conditions and the equipment that is installed [25]. Now, the literature till date points out towards two major approaches to deal with the problem: modeling the probabilistic behavior of the occupants and modeling their deterministic behaviors. However, over the years better predictions have been

attributed to the probabilistic modeling which logically makes sense, since the use of electricity in a residential building is completely dependent on the occupants and that behavior tends to be random, so it is important to incorporate that randomness in the model [26]. Going further, the probabilistic models can cover the factors like variability and diversity in an occupant's behavior by using empirical statistical data on which we estimate the probability of when a certain event will occur.

We utilize probabilistic modeling through machine learning to compare how the addition of socio-economic parameters of the occupants, to the regular time-series energy consumption data can enhance the prediction results [27][28]. The finalized load-profiles define the device-wise consumption pattern of different households. Based on the randomness of occupant behavior we discover how are the multiple appliances in a household being used, which combinations do the occupants prefer, what is the usage like during the peak load hours [29] and what we can infer from the predicted load profiles. For this purpose, we have chosen specific datasets that incorporate not only the total electricity consumption but the device-wise consumption as well. And provides the relevant socio-economic parameters like the numbers of occupants, size of the household, number of rooms, number of appliances.

#### **1.4. Research Objectives**

Following are the major objectives of this study

1. Building electricity load profiles based on device-wise consumption using the time-series data only.
2. Building electricity load profiles based on device-wise consumption using the time-series data and integrating socio-economic parameters with it.
3. Comparing, how the socio-economic parameters can contribute towards making better predictions.

#### **1.5. Overview of the Proposed Approach**

To accomplish the above-mentioned research objectives, we utilize probabilistic modeling through machine learning. First step is to generate the electricity load profiles only based on the historic data and then compare how the addition of socio-economic parameters of the occupants [30], to the regular historic, time-series energy consumption data can enhance the prediction results. The finalized load-profiles define the device-wise consumption pattern of different households. Based

on the randomness of occupant behavior, we discover how the multiple appliances in a household are being used, which combinations do the occupants prefer, what is the usage like during the peak load hours and what we can infer from the predicted load profiles. For this purpose, we have chosen specific datasets that incorporate not only the total electricity consumption, but the device-wise consumption as well, and provides the relevant socio-economic parameters like the numbers of occupants, size of the household, number of rooms, and number of appliances.

## **1.6. Structure of Thesis**

The rest of this thesis is organized as follows:

- Chapter 1: Research problem is introduced in the beginning of this section followed by main contributions of this work.
- Chapter 2: of the thesis presents detailed literature review. Analysis of other approaches from different domains to solve the similar research problem is summarized and discussed with shortcomings of other approaches. It will be shown with the literature review chapter that proposed work has not been done before and it will be useful to work in this area with proposed methodology.
- Chapter 3: Presents the methodology of the proposed Household Electricity Load Profile Generation System. The methodology contains the description of the detailed analysis performed on all the selected datasets, the final granularities selected, how the relation with socio-economic parameters was built. And finally, brief overview of machine learning models is presented.
- Chapter 4: Contains the results of this research work. The results are computed for individual appliances selected, their comparisons for both historic and historic plus socio-economic parameters are presented.
- Chapter 5: Presents the discussion of results, comparison with related techniques and limitations of proposed approach.
- Chapter 6: Presents the conclusion of the research and possible dimensions of future work.

## **Chapter 2**

### **Literature Review**

#### **2.1. Chapter Overview**

In this chapter, we discuss the related literature to highlight the work that has already been contributed towards the research in electricity load profile generation. The discussion of the previous literature helps us in justifying the proposed solution, pointing out the gaps that our research can fill.

#### **2.2. Background**

We have classified the electricity load profile generation systems based on the data sources that are used to generate the load profiles [31]. The literature being discussed classifies electricity load profiles that are generated using the historic data and those generated using socio-economic parameters along with the historic data. The three major techniques applied when using both the data sources are Machine Learning, Deep Learning and Data Mining based techniques.

##### **2.2.1. Electric Load Profiles Based on Historic Data**

Time domain analysis is one of the most popular methods of generating electricity load profiles using machine or deep learning techniques. The paper, “Characterizing patterns and variability of building electric load profiles in time and frequency domains”, highlights how buildings are a major electricity consumer, leaving behind a prominent carbon footprint [24]. And, how Building Energy Model (BEM) can play a significant part in designing and operating buildings that are energy efficient, helping in maintaining predictive controls which in turn helps in energy system planning. They initially signify the importance of advanced metering infrastructure as it leads towards the generation of a new data source, that can help in building electric load profiles at a higher temporal resolution. The authors point towards brighter side of improved electricity load profiles is that they can contribute towards an enhanced building energy model (BEM). The major advantage of identifying these characteristics is that they can assist in detecting changes in the electricity demand of a building [32], which could’ve occurred due to operational issues or some faults in the devices.

The paper focuses on using the new data source to propose a two-path approach, i.e., time-domain analysis and frequency domain analysis, to analyze building electricity load profiles with high temporal resolution. The former and more commonly used approach, i.e., time domain analysis can help us in extracting the core parameters that contribute towards characterizing the shape of the load, for example, the peaking load-ratio and the high-rise time. While, on the other hand, the frequency domain analysis, covers the part of identifying the major periodic electricity fluctuations and measuring the overall load variability.

To work around the proposed idea, the authors implemented and evaluated both the forth mentioned approaches, using 1 whole year's data that was recorded through a smart meter at an interval of 15 minutes. This was the data of 188 commercial buildings in North California. The results of both the approaches as per the results remain consistent and complement each other while representing full dynamics of the load profile. The final conclusions drawn from this study are that these analyses enhance the BEM's performance by providing highly realistic building operation plans and utilizing the developed variability metrics to match the simulated electricity load profiles against the real ones.

The paper "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data", kicks off by claiming how the development in the acquisition of the smart meter data has opened a path to new research in the electricity sector by making a large amount of real-time electricity consumption data available [33]. This data helps in electricity sustainability by finding the patterns of electricity usage by buildings, which can assist in improving building energy management. However, the authors point out that majority of the previous studies have focused on using this data to generate electricity usage patterns but have kept the work limited in terms of finding the hidden insights and applications of the pattern.

To break the limitation, the authors have proposed a framework that uses data mining techniques to find the typical electricity load patterns (TELPs) and then use these patterns to unfold the hidden insights. The framework is pipelined with three major tasks, i.e., data preparation, using various data mining techniques to find the electricity load pattern and finding information from within the discovered pattern. The authors have proposed an advanced clustering approach, consisting of a two-step clustering analysis to discover TELPs for individual buildings. To keep the

dimensionality of the load profile reduced, before clustering 5 statistical parameters are chosen that represents the shape of the load profile. In the two-part clustering, the first part is to detect the outliers from the daily electricity load profiles (DELPS) by using Density-based Spatial Clustering Application with Noise (DBSCAN). This clustering algorithm helps in addressing the quality issues in the electricity consumption data, that may have been risen because of the energy consumption monitoring platforms. The second phase of the approach enables the grouping of all the generated DELPS so that TELPS can be extracted, it uses k-means algorithm for the purpose. To verify the authenticity of the two-phase clustering a comparison is made with two single-step clustering approaches. Moving further, to gain insights from the extracted pattern Classification and Regression Tree (CART) algorithm is used. It also assists in improving the understanding gained from the clustering technique. The data used to analyze the performance of the framework is the time-series electricity consumption data of three commercial buildings in Chongqing, China. The results have shown an effectiveness and the extracted knowledge from the pattern is believed to help with early fault detection of anomalous electric load profiles. This framework can act as an electricity management framework for building managers providing them with understanding related to the usage pattern and the anomalies in it.

The paper “Applying load profiles propagation to machine learning based electrical energy forecasting” emphasizes on how electricity production is both an environmental and an economical challenge and that an optimal control on its production is the need of the hour [34]. To develop efficient forecasting systems [35][36], it is crucial that we must model the electrical energy correctly. To meet this need, authors introduce a novel approach to load forecasting that utilizes the load profiles (LPs). The study operates around Algeria, the first thing the paper covers is to analyze the power consumption in Algeria so they can learn about the various factors that effect it. Following the analysis, the hourly temperature profiles are used to apply the seasonal fluctuations. They use annual, weekly, and daily as the three levels of load-profile propagation to perform load-profile based forecasting. Artificial intelligence is used as the base method by the authors for the forecasting. They have utilized multiple AI techniques for developing both short and mid-term forecasting models. They have pioneered in using a two-dimensional Convolutional Neural Network (CNN) for the purpose of load foresting. The results the authors were able to get using the artificial intelligence (AI)based technique 2-D CNN for the load-profile based model was significantly high with MAPE being equal to 0.80%, RMSE being 75.57 MW, and the

Willmott's Index (WI) = 0.99. The results for the load profile propagation were MAPE being equal to 3.86%, RMSE being 372.68 MW and WI= 0.95. Comparison clearly shows that load-profile based AI model has higher tendency to produce better results.

The paper, "How to model European electricity load profiles using artificial neural networks" targets on generating synthetic yearly electricity load profiles for multiple European countries by utilizing the weather data, along with the electricity consumption data which is considered at a granularity of one hour [37]. To do so, they have utilized artificial neural networks (ANN), so that the long-term forecasting can be accomplished. For the training purpose they have utilized the historic electricity consumption data of Germany ranging from years 2006 to 2015. The paper majorly focuses on the utility of machine learning and how it can improve the generation of electricity load profiles.

The ANN is structured in a way that it has 5 hidden layers and 1024 hidden nodes on each layer. The key parameters fed to the model include the historic data that is stored in a calendrical manner, the peak loads occurring annually and the weather data. The proposed model is evaluated by comparing it against the current best model for generating synthetic electricity load profiles published by European Network of Transmission System Operators (entso-e). The year 2016 was selected to make the comparisons and the comparison metric was mean absolute percentage error, the proposed model took a precedence on the state-of-the-art approach by scoring an error of 2.8% as compared to 4.8%. Later, they have generated forecasts for Germany, Spain, France, and Sweden by utilizing the synthetic load forecasts that they have generated for the year 2025.

Another interesting fact is that they have used their proposed research to highlight the importance that the external temperature has on the predicted load profile. The authors believe that one of the major uses of these electric load profiles is that they will enhance the overall prediction accuracies of the electricity forecasts.

The paper, "Watt's up at Home? Smart Meter Data Analytics from a Consumer-Centric Perspective" discusses how smart meters have the advantage of transferring consumption data to remote computer systems rather than traditional metering devices, which gives us the benefit of collecting a new dataset [16]. The data collected by these devices does not only help in the calculation of a customer's electricity bill, but also serves for a variety of novel purposes. While many of these services aim to improve the overall operation of the power grid, most of them are



tailored specifically to address these needs. A prominent use case would be how forecasting the energy consumption of a household, or the photovoltaic production can help in improving the power generation schedule of a power-grid. Analyzing the consumption patterns can lead towards detection of the anomalies which can also serve as indicators of electricity theft and warrant investigation. The authors emphasize on the fact that even though electricity consumption is something that is completely dependent on the user pattern but, the research in the sector usually benefits the grid more generally than the user. Keeping these stats in mind the authors use this research to review the range of services that can be used to benefit the end-users. Their research focuses on exploring the state-of-the-art methods and targets the data communication and processing gaps so that this research can be molded into becoming a prospective solution of a consumer-centric electricity problem. The major hurdles while designing a consumer-centric system are usually; lack of standardization in the data, algorithms that provide mediocre results and the privacy concerns, the authors believe that advancement in these ideas can lead towards a huge development shift.

The paper “Day-Ahead Short-Term Load Forecasting for Holidays Based on Modification of Similar Days' Load Profiles” emphasizes on the importance of Short-Term Load Forecasting (STLF) and how it is getting harder every day with the introduction of more and more distributed resources in the power system [38]. The most common example of these distributed resources are behind-the-meter (BTM) PV resources that are added to the power systems. The authors in this study, majorly focus on overcoming the issues that are faced while performing short-term load forecasting by creating a framework for STLF for holidays only. For this, they have considered four major factors i.e., datetime (calendar), weather, trend, and the BTM-PV (which is believed to be a major reason for the disturbance in the STLF values), these factors play a prominent role in the overall generation of the electricity load profiles.

The first thing in the framework is to pair the holiday that is being targeted with all its corresponding days in the history, this enables the addition of the calendar factor, all these days that are paired together are labeled as **similar days**. Moving on, to incorporate the remaining three factors the difference of days between the target holiday and the historic holidays is computed and the effect that they reflect on the load difference (difference with the factors induced) is then

quantified. The next process is to generate the load profile on each of the generated pairs and once the load profiles are generated, they are then combined for the entire holiday.

To test the working of this framework, this solution was implemented on Korean National Holiday case study, the output of this test was compared with the performance of the conventional model. The output results clearly point out how the proposed framework outperforms the conventional system. The results point out that this model can be implemented for the STLF of the holidays and can help in improving the overall accuracy of the forecasting.

### **2.2.2. Electricity Load Profiles Based on Socio-Economic Parameters**

In paper “Estimating hourly lighting load profiles of rural households in East Africa applying a data-driven characterization of occupant behavior and lighting devices ownership” the authors develop a way of estimating the electricity consumption patterns in rural East-Africa where, due to increased inflation and lack of resources people mostly rely on electricity only for using the light-based devices, through literature it was found that in an average East-African household more than 50% of the total electricity consumption is credited to the lightening devices [39]. Now, developing a model to fetch the electricity consumption pattern in a rural area is a tedious task because of the lack of data availability and the model complexity. Since, based on such small data computing the load profiles can be challenging with increased chances of error-prone results.

The research in the current paper is built around these problems, to tackle them through a perspective solution. Since the easy availability is only of the lightening data, so using this small data as input, the authors have developed a methodology in which they generate the hourly lightening load profiles of the rural households in East-Africa. For the preprocessing of the data, to enhance the overall result quality, the authors have integrated weather data and the satellite imagery with the household lightening data. After the integration, they have used machine learning as part of the methodology to predict the behavior of the occupants based on their indoor-outdoor lamp usage. Based on this approach, the average prediction accuracy acquired by the authors is 80%. 13 households in Kenya have been measured manually, so that validation data is available to validate the performance of the model once the light functions have been applied. The overall testing of the model shows that it can generate the rural household electricity consumption load profiles with an average normal root mean squared error of 0.7%, which is less as compared to the error values obtained through simulation-based approaches that use the on-site data for predictions.

To conclude the research and show a demonstration of the application of the research, the authors have computed 1 month's load profiles of the household and projected them over the households in Kenya to explore the real-time outcomes.

The paper "Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network" points towards the importance of electricity load forecasting by stating how they contribute towards better energy management, total budgeting and in building an improved infrastructure, this is the reason why advanced research is being carried out in the sector [17]. Over the decades, as we witness a blooming increase in the installments of smart meters and other consumption recording sensors in buildings and even in individual households, we get an opportunity to make the sensor-based forecasting possible.

The key focus of the authors is to utilize the deep learning architecture to accurately utilize the sensor data and make the forecasting. According to literature, the RNNs (Recurrent Neural Networks) are highly popular and accurate for the forecasting purpose however, one set-back of these models are that they are trained offline. Their learning is done based on the pre-collected data, losing the opportunity to train on the freshly arriving data. Another major issue is that the RNNs are not made to handle the concept drift, this can cause an overall negative effect on the forecasting for example, if the load has shifted due to the installation of a new device. If these issues are tackled, RNN becomes one of the best deep learning architectures for forecasting. This paper majorly focuses on targeting these issues, by proposing an Online-Adaptive RNN. As the name implies, this architecture incorporates the online learning within the traditional RNN, making it capable of learning new data as it arrives and changing along with the new patterns.

The authors have modified the traditional RNN such that they utilize it for the time dependency factor and to incorporate the online learning they have made changes such that the model weights are updated as per the new incoming data. The model performance is continuously monitored, if at any point the accuracy starts to face degradation, the online training is activated and the model starts learning on the incoming data, hence upgrading the hyperparameters to meet the new need. To test the performance of the model, the authors have considered 5 households, the results clearly point out an improved forecasting accuracy as compared to the offline RNN and the regular long-short term memory networks. It also takes precedence on 5 other online-training models. In terms of training time, another interesting and winning fact about the proposed model is that online

training only takes fraction of the time that the model would take when it is trained explicitly offline.

The paper “Day-Ahead Short-Term Load Forecasting for Holidays Based on Modification of Similar Days' Load Profiles” emphasizes on the importance of Short-Term Load Forecasting (STLF) and how it is getting harder every day with the introduction of more and more distributed resources in the power system [38]. The most common example of these distributed resources are behind-the-meter (BTM) PV resources that are added to the power systems. The authors in this study, majorly focus on overcoming the issues that are faced while performing short-term load forecasting by creating a framework for STLF for holidays only. For this, they have considered four major factors i.e., datetime (calendar), weather, trend, and the BTM-PV (which is believed to be a major reason for the disturbance in the STLF values), these factors play a prominent role in the overall generation of the electricity load profiles.

The first thing in the framework is to pair the holiday that is being targeted with all its corresponding days in the history, this enables the addition of the calendar factor, all these days that are paired together are labeled as similar days. Moving on, to incorporate the remaining three factors the difference of days between the target holiday and the historic holidays is computed and the effect that they reflect on the load difference (difference with the factors induced) is then quantified. The next process is to generate the load profile on each of the generated pairs and once the load profiles are generated, they are then combined for the entire holiday.

To test the working of this framework, this solution was implemented on Korean National Holiday case study, the output of this test was compared with the performance of the conventional model. The output results clearly point out how the proposed framework outperforms the conventional system. The results point out that this model can be implemented for the STLF of the holidays and can help in improving the overall accuracy of the forecasting.

The paper, “Modeling and analysis of the electricity consumption profile of the residential sector in Spain” focus on generating electricity load profiles for domestic households in Spain [14]. They point out, how detecting the pattern of electricity consumption in residential sector is an extremely tedious task, given the variations in the behavior of the residential consumers. Residential sectors come with a lot of diversity that is reflected in terms of the household sizes, the electronic devices, location of the house and most importantly the variable behaviors of the residents in the household,

who define the consumption patterns. Considering all these factors, computing the electricity consumption profiles in residential sector is a highly expensive task.

To avoid the high costs, in this paper, a simulation model is created that utilizes a bottom-up stochastic approach to compute the electricity load profiles in residential sector, using the data provided by Survey of Time Employment of the National Institute of Statistics of Spain (INE). The simulated algorithm generates an average electricity consumption load profile based on two factors: number of people residing in the household and the current day of the week.

The results gained through the simulated algorithm are all kept separately for each household for the sake of being analyzed separately. The authors have generated electricity load profiles so they can be utilized as a baseline for initiating multiple other research tracks. They specifically use these results to focus on self-consumption, how individual households can improve or stabilize their consumption behavior based on the consumption pattern of their house [40], [41]. Other than this, the study can become a perspective baseline for targeting energy-efficiency problems, demand-side management and for computing new energy policies.

The paper “Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN)”, explains how the electricity load profile of a building can provide a fulfilling view of the electric utilization in that building [42]. They also provide a thorough understanding of energy efficiency in a building and what is the measure of demand side flexibility. The behavior of a real-time building is both dynamic and stochastic and the current approach to generating the electric load profiles is slow and unable to capture this behavior properly. Another major issue is that some of the approaches used to generate the load profiles cross the line of personal security.

The authors have relied purely on machine learning to generate realistic electricity load profiles, while attempting to avoid the issues mentioned before. They have utilized Generative Adversarial Networks (GANs), which is a powerful machine learning algorithm, capable of extracting the hidden probability distributions by using the data alone. The author’s proposed work is composed of three major steps; bringing the daily 24-hour load profiles in a normalized form, then using the K-means algorithm they cluster all the load-profiles. Finally, they apply GANs on each cluster to generate new daily load-profiles for each of the cluster. To test the pipeline, they have used the Building Data Genome Project data, which is an open-source database available. For the validation

purpose, the authors have made a comparison between the new and the real load profiles by computing the differences in mean, standard deviation, and the overall distribution of the key parameters of both load profiles. They have also utilized KL divergence and discovered that majority parameters of the new and real load profiles are within 0.3.

Furthermore, the GANs proved helpful because they don't only focus on capturing the general pattern but also the slight random variations that occur in a building's load consumption. Overall, this approach is useful for generating the electric load profiles, verifying the other models used for generating load profiles, capture changes within load profiles and lastly it has ability to do all the above in an incognito manner, hence, avoiding any privacy breaches to the users. Thus, it enables a support for the research targeting energy efficiency.

### 2.3. Conclusion and Comparative Summary

The primary objective of this chapter is to explore, shortlist, and study the published research work in the domain of electricity load profile generation. The literature review discusses different tools and techniques proposed through research for better and efficient predictions. Table 2.1 summarizes the comparison of state-of-the-art research. Research community has given great attention to the area of generating electricity load profiles for buildings in the past few years, however, it needs to be more closely observed in terms of individual devices by combining both historic electricity consumption data and the socio-economic parameters.

Table 2.1: Side by side comparison of different electricity load profile generation strategies in the literature

	<b>Objective</b>	<b>Historic Data</b>	<b>Socio-Economic Data</b>	<b>Device-Wise Segregated Data</b>	<b>Technique</b>
[24]	Building electricity load profiles with high temporal resolution	Yes	No	No	Time and Frequency Domain Analysis
[33]	Generate a framework that can extract typical electricity load patterns for more useful insights	Yes	No	No	DBSCAN, Classification and Regression Tree

[34]	Effective load-forecasting approach using load profiles	Yes	No	No	2D-Convolutional Neural Nets
[37]	Generating Annual electricity load profiles using AI	Yes	No	No	Artificial Neural Nets
[16]	Finding potential of smart-meter data from user centric point of view.	Yes	No	No	
[39]	Designing solutions for enhanced energy access in rural areas.	Yes	Yes	Yes (only lighting devices)	Cluster Analysis, Random Forest
[17]	Improving RNNs for improved load forecasting	Yes	No	No	Online Adaptive RNN
[38]	Framework that allows short-term load forecasting for consumption behaviors on holidays	Yes	No	No	Short Term Load Forecasting
[14]	Creating average load-profiles for predicting daily electricity consumption patterns	Yes	Yes	No	Simulation with Bottom-up Stochastic Approach
[42]	Generating synthetic electricity load profiles using GANs	Yes	Yes	No	K-means, GANs

## Chapter 3

### Methodology

In this chapter, we discuss the strategies that we have adapted to formulate the solution of our given research problem i.e., building electricity load profiles for residential buildings. Major focus will be on the data that we considered and used for the simulated solution, the analysis and the technicalities that have been used.

The main target of the study is to generate electricity load profiles with improved prediction scores. To do this, we have defined the research objectives, as follows:

1. Building electricity load profiles based on device-wise consumption using the historic data only.
2. Building electricity load profiles based on device-wise consumption using the historic data and integrating socio-economic parameters with it.
3. Comparing, how the socio-economic parameters can contribute towards making better predictions.

Total energy consumption readings of a household are not enough to identify the pattern of how the residents within the accommodation are utilizing the energy resources and similarly the prediction values of total energy consumption don't make much sense.

The target of this research is to predict how the electricity is being consumed by the individual devices in a household. Our approach generates a pattern, an understanding of how residents are consuming electricity. For example, are the residents in a specific household using the washing machine and the air conditioning at the same time or are they using one at a time. The answers to such questions are further enhanced with the addition of socio-economic parameters. The size of the house, the number of rooms, number of people, answers why and when a specific device would be turned on or off.

For generating a solution, we selected 4 different datasets, the first step was to analyze them in their raw form, to uncover any underlying insights, then we pre-processed the data and performed a post analysis. Once this is complete, we move towards the training of the data for which we have selected two different machine learning models and one deep learning model, for the sake of



comparison. We have selected multiple evaluation metrics to determine the performance of our models. Figure 3.1 provides a graphical representation of the pipeline used to generate household electricity load profiles.

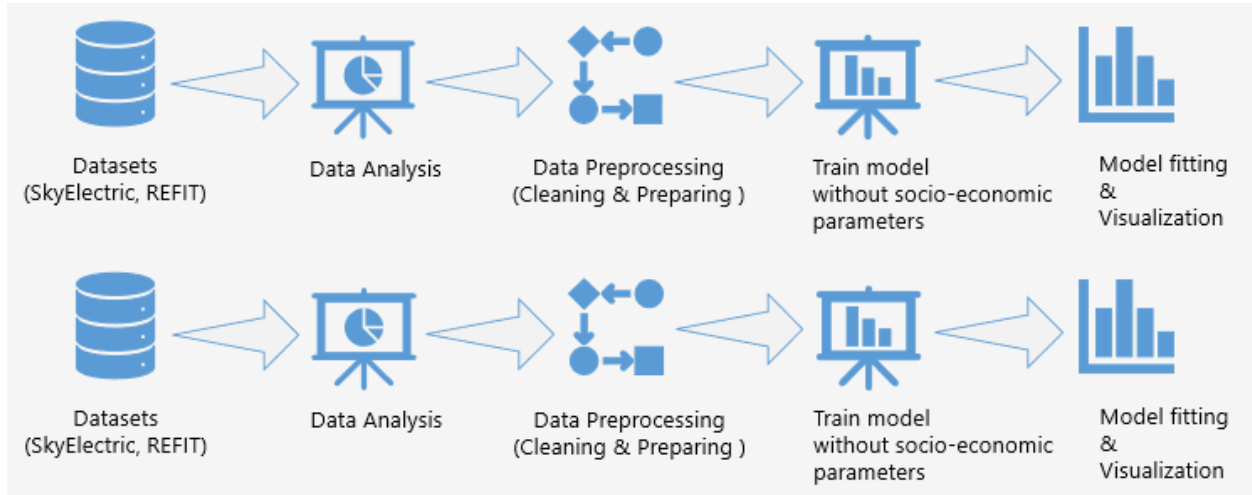


Figure 3.1: Pipeline to Generate Household electricity Load Profiles

All the steps that serve as part of the methodology have been discussed in detail from here onward.

### 3.1. Datasets

The building unit of this research was the electricity consumption data of residential buildings. Though there are multiple publicly available datasets that could aid the research being done in the electricity sector but, the specificities that we require from the data had limited us to a few. We initially began with 4 different datasets

1. REDD (Reference Energy Disaggregation Dataset) [43]
2. PRECON (Pakistan Residential Electricity Consumption Data) [22]
3. REFIT Electrical Load Measurements Dataset [44]
4. Sky Electric Dataset (Publicly un-available)

We choose a wider range of datasets to understand the underlying patterns that exists in electricity consumption data, how we can analyze and pre-process it in a way that we get the maximum information from it. Following is brief description of each dataset and the analysis performed under this research.

### **3.1.1. REDD**

Reference Energy Disaggregation Dataset [43] was the first data that we started processing. It is a large scale publicly available dataset that contains aggregated power usage data and the disaggregated data that is collected from each individual circuit of multiple households. The data is recorded for refrigerator, kettle, microwave, television, etc. The purpose of designing this dataset is to tackle the energy sustainability issues with assistance of machine and deep learning models. The task in mind while developing the dataset was to disaggregate the cumulative power usage of households into the device-wise consumption. But we utilize the data to understand how the historic data of a household can unveil important patterns presented by the electricity consumers.

The dataset has been recorded based on two main types of electricity data:

- High Frequency, current or voltage data from the two power mains.
- Low frequency, current or voltage data which include data of both the mains and the individual circuits.

An in-detail view of the dataset shows that the data has been recorded for 10 houses over a period of 119 days. For every monitored house the data has been collected for the electrical signal of the entire house at a frequency of 15kHz (adding current monitors on both power phases and voltage monitor on one phase), about 24 different individual devices in a household have been recorded; within the dataset they have been labeled by the name of the device they have been collected from. These devices have been recorded at a frequency of 0.5 Hz. 20 plug level monitors have also been recorded from each house at a frequency of 1 Hz, all these plug level logging devices have been grouped under a single circuit. After combining all these circuits there were a total of 268 monitoring devices that enabled the generation of 1 Tb of raw data during the research period.

#### **3.1.1.1. Analysis of REDD**

We consider 6 houses from the dataset, with the electricity consumption duration recorded from 2011-04 to 2011-05. For each house, the individual devices that have been recorded are different, hence the visualizations are different for each household. To get a clear understanding of the common everyday devices, we use the data from the first house and visualize the devices in an independent way.

### 3.1.1.1. Identification of Cold Appliances

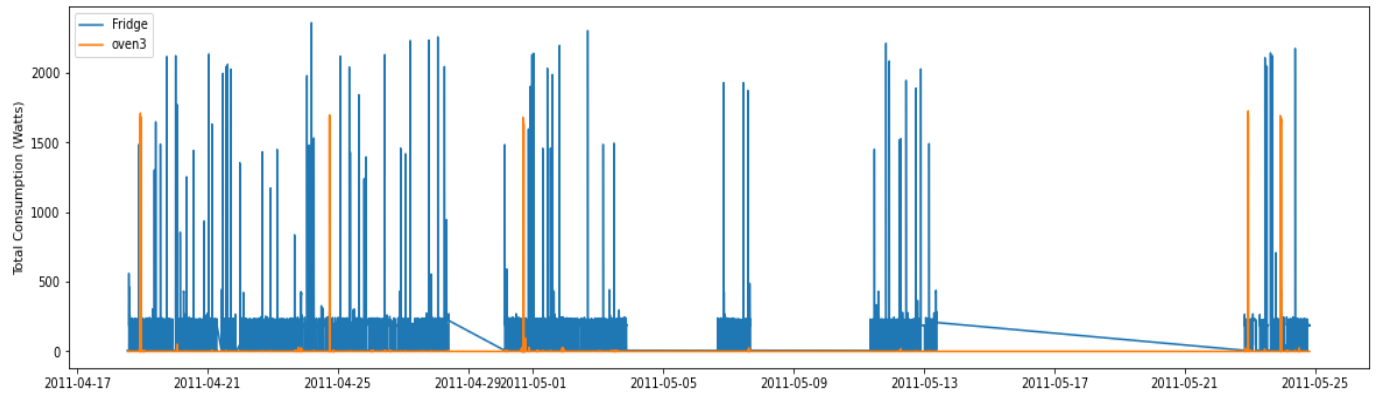


Figure 3.2: REDD House 1; Refrigerator Consumption Pattern with Respect to Oven

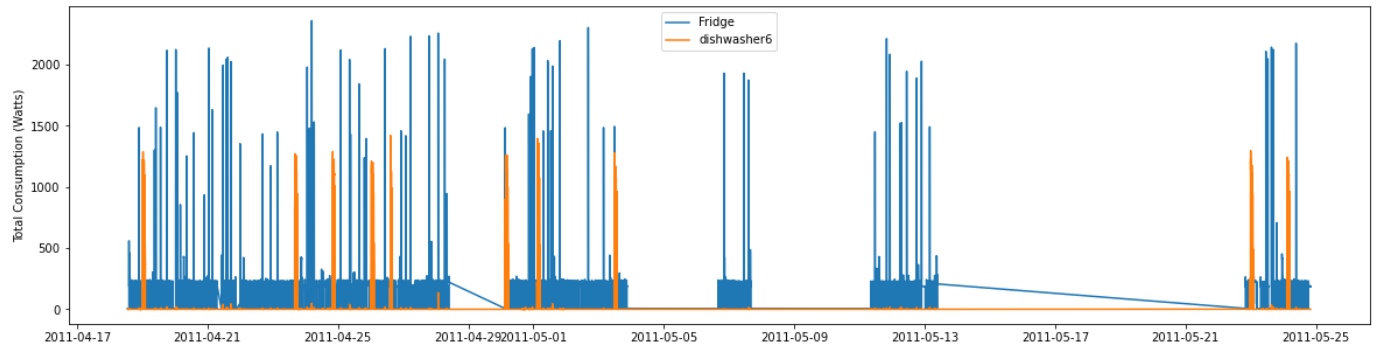


Figure 3.3: REDD House 1; Refrigerator Consumption Pattern with Respect to Dishwasher

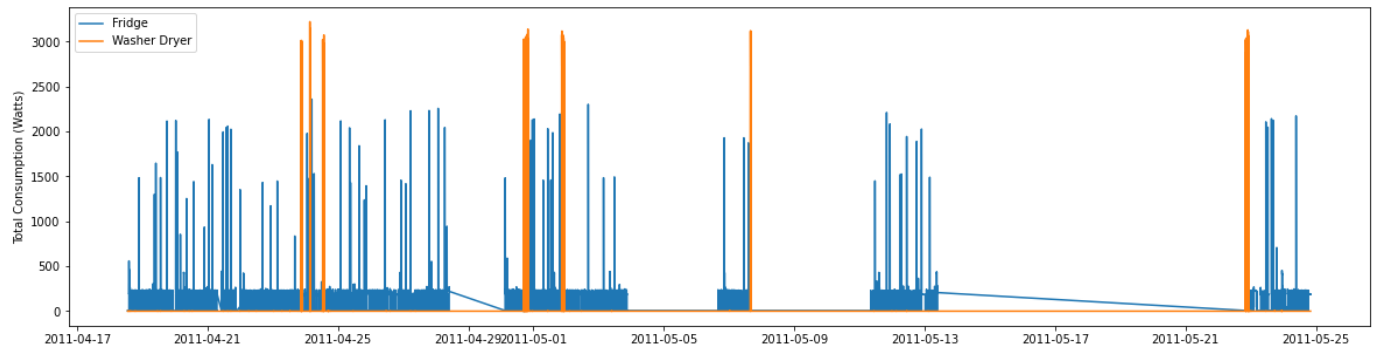


Figure 3.4: REDD House 1; Refrigerator Consumption Pattern with Respect to Washer Dryer

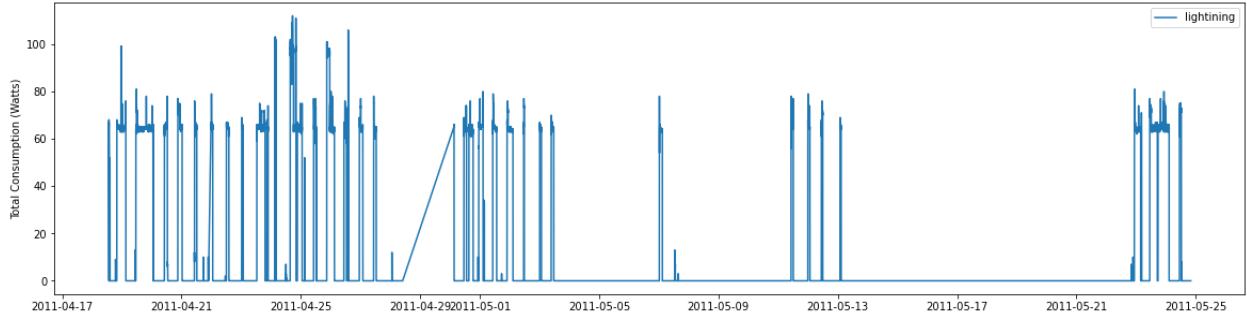


Figure 3.5: REDD House 1; Lighting Device Consumption Pattern

The first three figures point towards constant consumption of electricity by the refrigerator as we compare it against multiple high voltage electricity consumption devices. And the last figure points towards the consumption pattern of the lighting device. This allows us in categorizing the refrigerator and lighting devices as a cold appliance which are in a continuous state of consumption. We have plotted separate graphs because the consumption by power (in terms of Watts) of lighting devices is very small as compared to the refrigerator and hence the visualization is not clear.

Since, we do not consider the cold appliances as part of this research, now, we see an overall pattern of consumption but, we remove the cold appliances.

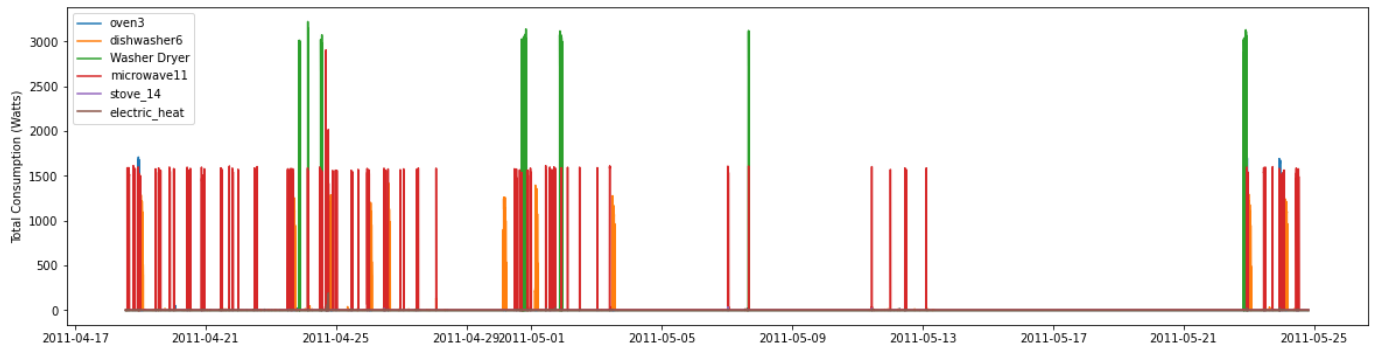


Figure 3.6: REDD House 1; Device-Wise Consumption Analysis

We notice that in current raw data visualization it is hard to deduce the electricity consumption load profiles.

To support the overall insights that we have gained from the analysis of house 1, we draw out a visualization for 4 houses, excluding all the cold appliances.

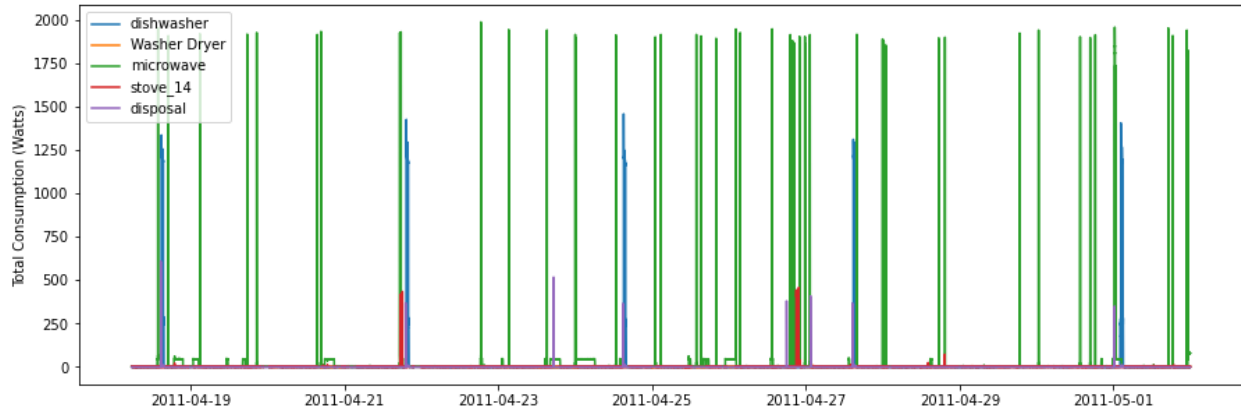


Figure 3.7: REDD House 2; Device-Wise Consumption Analysis

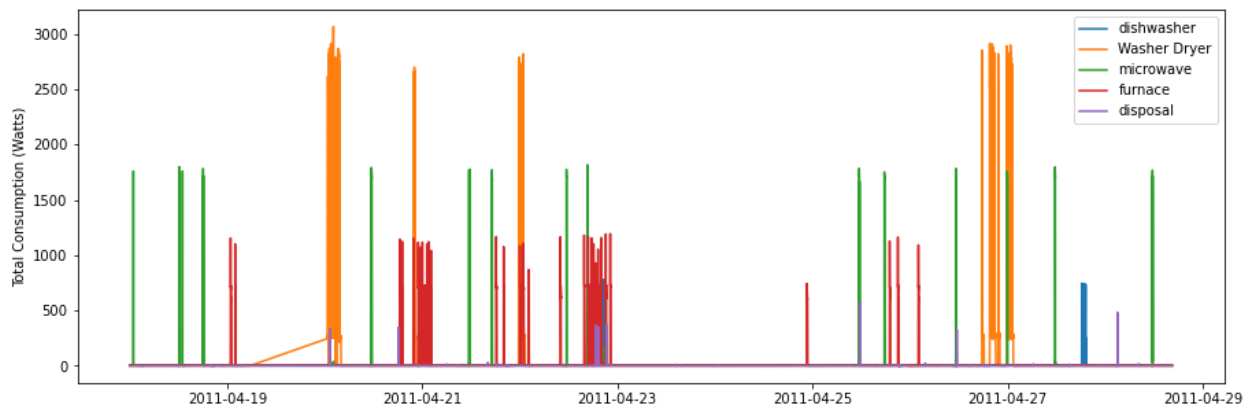


Figure 3.8: REDD House 3; Device-Wise Consumption Analysis

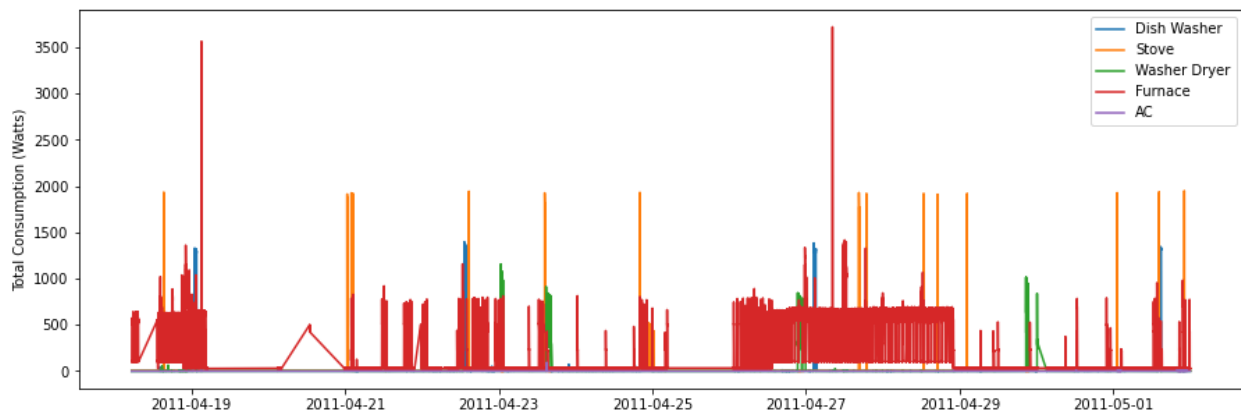


Figure 3.9: REDD House 4; Device-Wise Consumption Analysis

Each house has a different set of parameters (the monitored devices), but we can notice that there is some distinction between the usage of the devices. All the devices visualized above are high

voltage devices, and from visualization we can tell that device like microwave has been used standalone in most of the cases.

One of the major setbacks in REDD for us, was the inconsistency in the devices that were monitored from each house, to generate a load profile we require merging the data of all the houses so we can use it for the training purpose, for this we require at least 5 devices that are similar amongst all the houses. In addition, the data comes without the socio-economic parameters and the publicly available size of the data is rather limited. Hence, post this analysis we planned to proceed without considering REDD further for our current cause.

### 3.1.2. PRECON

Around 40% of the total electricity generated world-wide is consumed by the residential sector and hence, for taking energy efficient measures it is important to understand the behavior of the residing occupants. Data is a key for understanding the depth of a problem, but the region of South Asia and Pakistan in specific, has a very few, small publicly available datasets that are not enough to understand the underlying patterns of electricity consumption. PRECON (Pakistan’s Residential Electricity Consumption) is a dataset that has been specifically designed to target this issue [22].

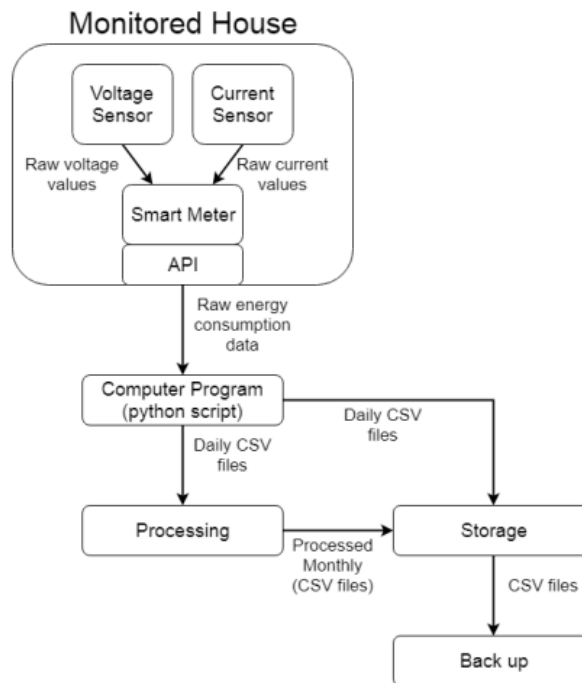


Figure 3.10: Data Collection Architecture for PRECON [22]

PRECON has been recorded for 42 Pakistani households at a granularity of 1 minute. The data has been recorded for the duration of 8 months, starting from 1<sup>st</sup> June 2018. Every 1440 rows in the data represent one single day, timestamp is used to keep the time record. The data is order in a way that the first column is always the timestamp, and the second column is always the total usage. The rest of the columns for each household differ, because the number of appliances vary from house to house.

Other than the electricity consumption data, this dataset also comes with the socio-economic details of the households. They have recorded the total number of people that are living in a household, this attribute is further divided into number of adults, children, and elderly that are the residents. The demographics of the households i.e., the size of the household, total number of rooms; this attribute is further divided into the number of bedrooms, kitchens, drawing and living rooms, number of floors, height of the ceiling, the year in which the house was built, and the number of each electronic device in the household are also recorded. All these attributes are believed to have a positive correlation with the electricity consumption and can significantly contribute towards generating better household load profiles.

#### **3.1.2.1. Analysis of PRECON**

PRECON is a Pakistan-based dataset, it allows us to get an in depth understanding of how the Pakistani residential community is consuming electricity. For PRECON, the initial analysis will be displayed for only one house. The reason is that all the houses share same characteristics, similar devices, though the socio-economic values and the consumption patterns are different however, the initial insights can be captured from a single house.

The analysis on REDD has already led us to confirm about the cold appliances and how they don't influence much on the household electricity load profiles. However, as a confirmation that the cold appliances perform similar in Pakistan, (separate ethnic region) we will display the electricity consumption of refrigerator. PRECON, does not record the data of individual lighting device so, we are unable to capture the pattern for them. From the Figure 3.11 we can observe that for the entire duration for which the data has been recorded the consumption of refrigerator is constant.

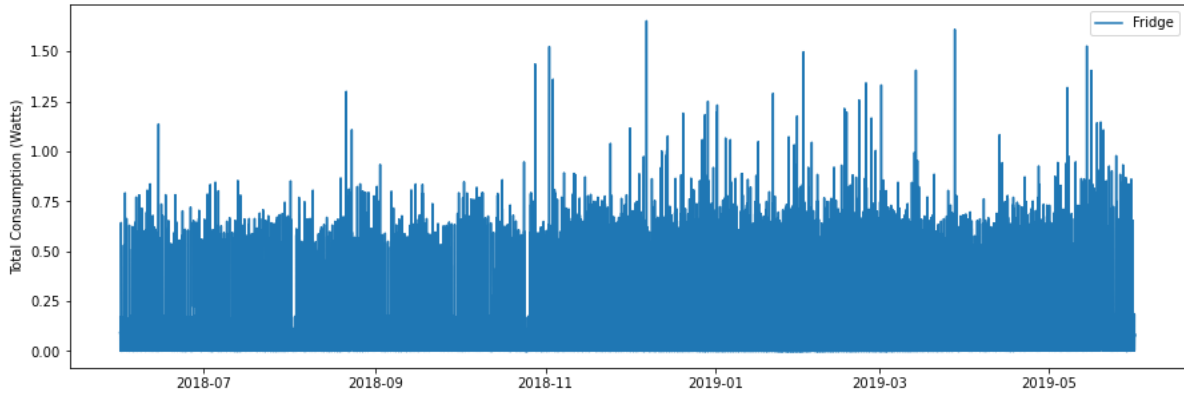


Figure 3.11: Total Consumption Pattern of Refrigerator

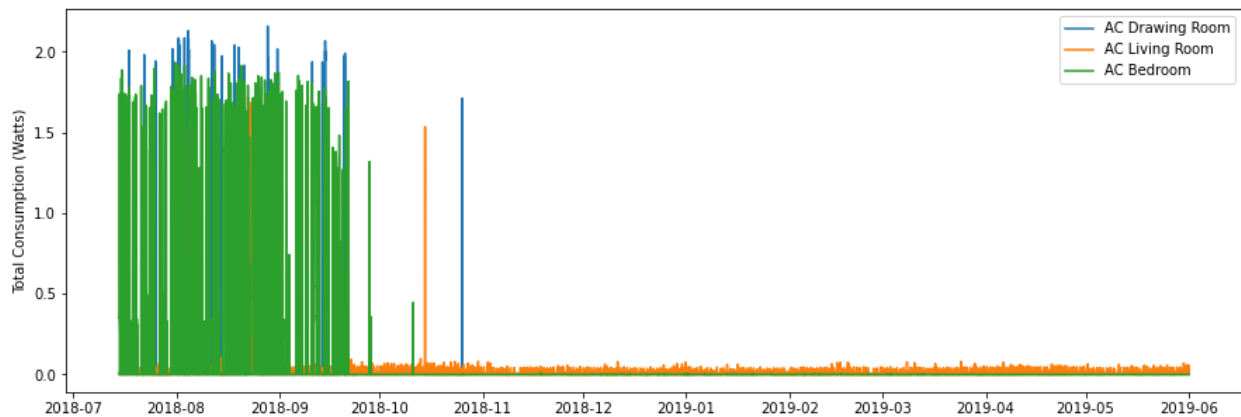


Figure 3.12: Total Consumption Pattern of the Individual Monitored Devices PRECON

Figure 3.12 is a visualization of the monitored devices without considering the cold devices, the data does not offer a wide variety of devices that we can play with, however, it provides a very clear picture of consumption. The recorded values are of 3 different air-conditioners in a household, and we can easily see that the consumption of all three devices is done in the hotter temperature months (July, August, September, and some part of October). This gives us the insight that the electricity consumption pattern is strongly reliant upon the outdoor weather.

This is a very broad picture, however, further ahead, we describe segmentation of the historic data based on different time windows and then establish a relationship between the socio-economic factors and the historical data.

### 3.1.3. REFIT

REFIT electrical load measurement data [44] is one of the biggest publicly available datasets that provides real-time electricity consumption data collected through smart meters from 20 different



houses in the UK over a period of two years. This dataset comprises of both historic and socio-economic data, and other than the total electricity consumption, it also provides the monitored data of the appliances within the households.

The data consists of 1,194,958,790 total readings and represents data that has been collected from over 250,000 appliances over a period of 2 years. The total electricity consumption reading and the consumption readings of the all the individual devices have been recorded at a granularity of 8s. A total of 9 appliances have been chosen and monitored from each household however, the devices are not the same but, it is to be noted that majority of the households share similar common devices for example, television, washing machine, dishwasher, microwave, etc. All the data has been recorded during active work time of the house residences and hence the data is referred as the real data.

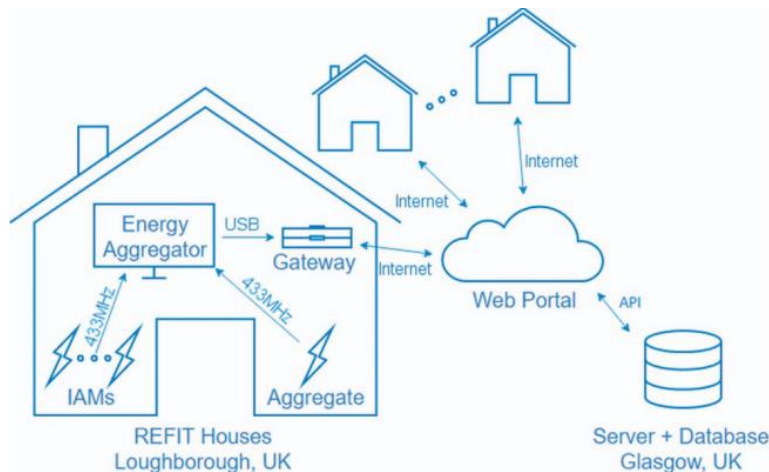


Figure 3.13: Schema of REFIT Data Collection [45]

Other than the electricity consumption data, REFIT contains socio-economic information of all the 20 households. It includes information regarding the number of occupants residing in the households, the estimated dwelling age of the house, the total number of appliances and the number of rooms in each household. All these attributes are positively correlated to the electricity consumption and contribute a great deal in the accurate predictions of the household load profiles.

### 3.1.3.1. Analysis of REFIT

By now we have thoroughly deduced the pattern of the cold appliances and hence, in the analysis of the REFIT we will no longer visualize such appliances. REFIT is one of the most thorough and sort out dataset and hence provides a wide range of insights. In the current visualizations, we

witness the consumption patterns of all the monitored devices within a single household that has been recorded.

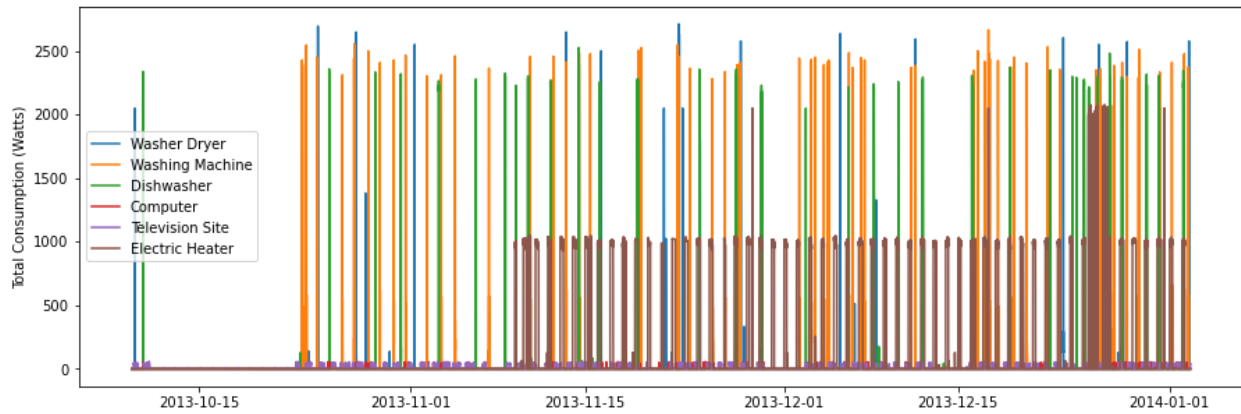


Figure 3.14: Total Consumption Pattern of the Individual Monitored Devices REFIT

Based on the Figure 3.14 we have seen that the total consumption pattern of REFIT based on the historic data does not reveal too many details, which is justifiable because all the monitored devices are majorly independent from the outdoor weather conditions. However, the interesting fact is that all the devices have a close connection with the socio-economic parameters. Segments of the data will be able to give a clearer vision of how exactly we can retrieve information from the REFIT data.

### 3.1.4. Sky Electric Dataset

Sky Electric dataset is a very small privately owned dataset, that we have used to specifically fetch the patterns of electricity consumption in the Pakistani Households. The data has been collected over a period of one month from a single household. The data has both electricity consumption readings of the entire house and of the individual appliances in the household. Data has been recorded at a very low granularity of just 1 second. The data does not come with socioeconomic parameters; however, it serves as test set to evaluate the performance of historic electricity consumption prediction model.

#### 3.1.4.1. Analysis of Sky Electric Dataset

Sky Electric is a small Pakistan-based private dataset that has limited number of data; however, we are utilizing the dataset as a test set to see how well we were able to predict device-wise consumption only based on the historic data. Like PRECON, this dataset also majorly comes with

monitored air-conditioning devices. Reason behind this is that till date, average Pakistani household doesn't over rely on electronic devices like electric stoves or dishwashers.

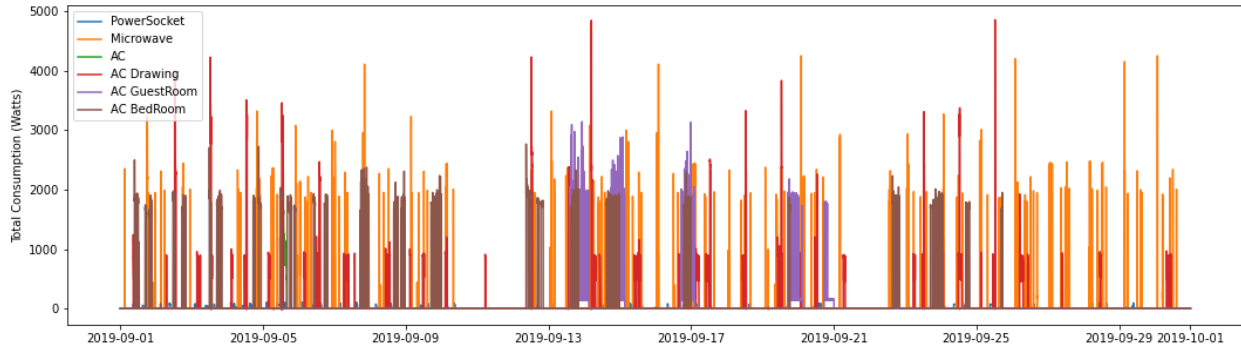


Figure 3.15: Total Consumption Pattern of the Individual Monitored Devices Sky Electric

Interesting fact to note is that since we only have data for the month of September, and September is a hot month the use of the air conditioning is quite frequent. This coincides with the insights that we collected from the PRECON data regarding the weather dependency.

### 3.1.5. Dataset Summary

After the required analysis of the data, we decided to go with only REFIT and Sky Electric for the purpose of this research. The reason is that both the datasets provide a complete picture of the device-wise electricity consumption in households, whereas PRECON fails to do so. The monitoring is limited to only air-conditioning devices in 85% of the households. It basically means that we have the same device to play around with an addition to the cold appliances.

Table 3.1: Summarized Comparison of the Datasets

Dataset	Size (Period over which data is recorded)	Number of Monitored Devices	Socio-Economic Parameters	Selected for Research
REDD	Less than 1 month	24	No	No
REFIT	2 years	9	Yes	Yes
PRECON	1 year	3 or 4	Yes	No
Sky Electric	1 month	10	No	Yes

## 3.2. Pre-Processing

After dataset selection the next important step is to mold the data in a way that it caters our needs. Through our visual analysis, we have already seen that raw data cannot help us much in the task of predicting the device-wise consumption of a household accurately.

The pre-processing steps are very straight forward and are mentioned as follows:

1. Finalizing Data Features
2. Finalizing Data Granularity
3. Altering Data Representation (Converting Data to Binarized Format)

### 3.2.1. Finalizing Data Features

In every dataset the total number of devices monitored are different, however, to create a stable dataset that is large enough to train the model we require similar features from every household to be concatenated together. For this purpose, we had to find the features (devices) that are common in majority of the houses in the REFIT Dataset. Generally, the most common devices were fridge and freezer, but they are cold device, so, we went on without them. After going through the devices in all the households and analyzing their usage, following devices as shown in Table 3.2 were selected from both the datasets.

Table 3.2: Selected Features from REFIT and Sky Electric

Datasets	Microwave	Power Socket	HV Bulb	AC	Washing Machine	Dishwasher	Television	Kettle
SkyElectric	✓	✓	✓	✓				
REFIT	✓				✓	✓	✓	✓

#### 3.3.1.1. Weekday as a Feature

Once the devices were finalized, the next important step was to add an enhanced impact of the date and time values that were provided with the data. Using the dates, we extract the day of the week from the data. We could extract the month since month of the year provides us the weather information, but, for both our datasets and the devices that we have been able to extract the outdoor

weather doesn't carry much effect. However, through visualizations, we can see that the day of the week has an interesting association with the electricity consumption.

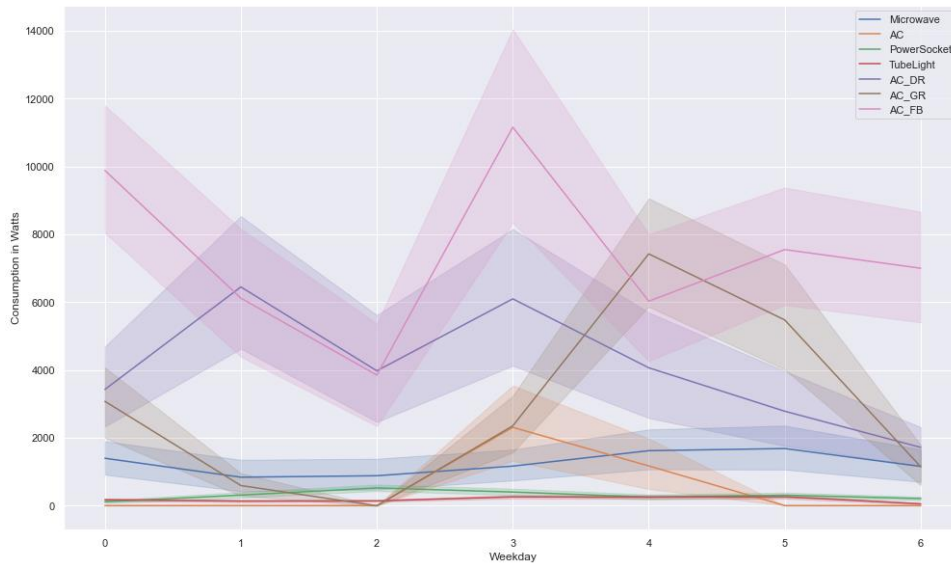


Figure 3.16: Impact of the weekday on the electricity consumption of the devices - Sky Electric

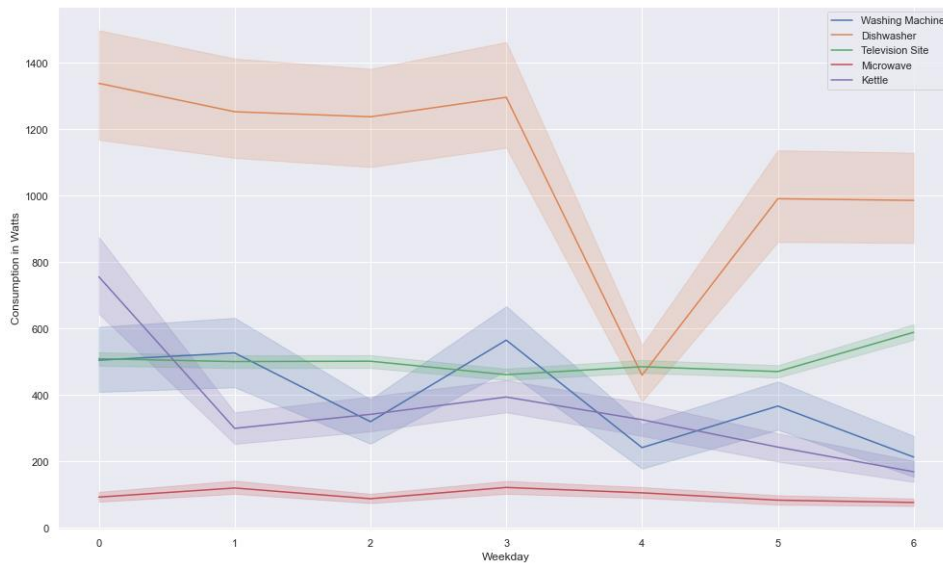


Figure 3.17: Impact of the weekday on the electricity consumption of the devices – REFIT

### 3.3.1.2. Merging Socio-Economic Features in REFIT

REFIT provides us with multiple socio-economic features which can influence the prediction accuracy of device-wise electricity consumption leading towards improved household electricity

load profiles. The socio-economic features that we have selected as part of the training data are as follows:

1. Occupancy, number of residents residing in the household.
2. Appliances Owned, number of total appliances owned.
3. Type, whether the household is detached, semi-detached or mid-terrace
4. Size, the number of rooms in the household.

We observe a correlation amongst all the finalized features for REFIT.

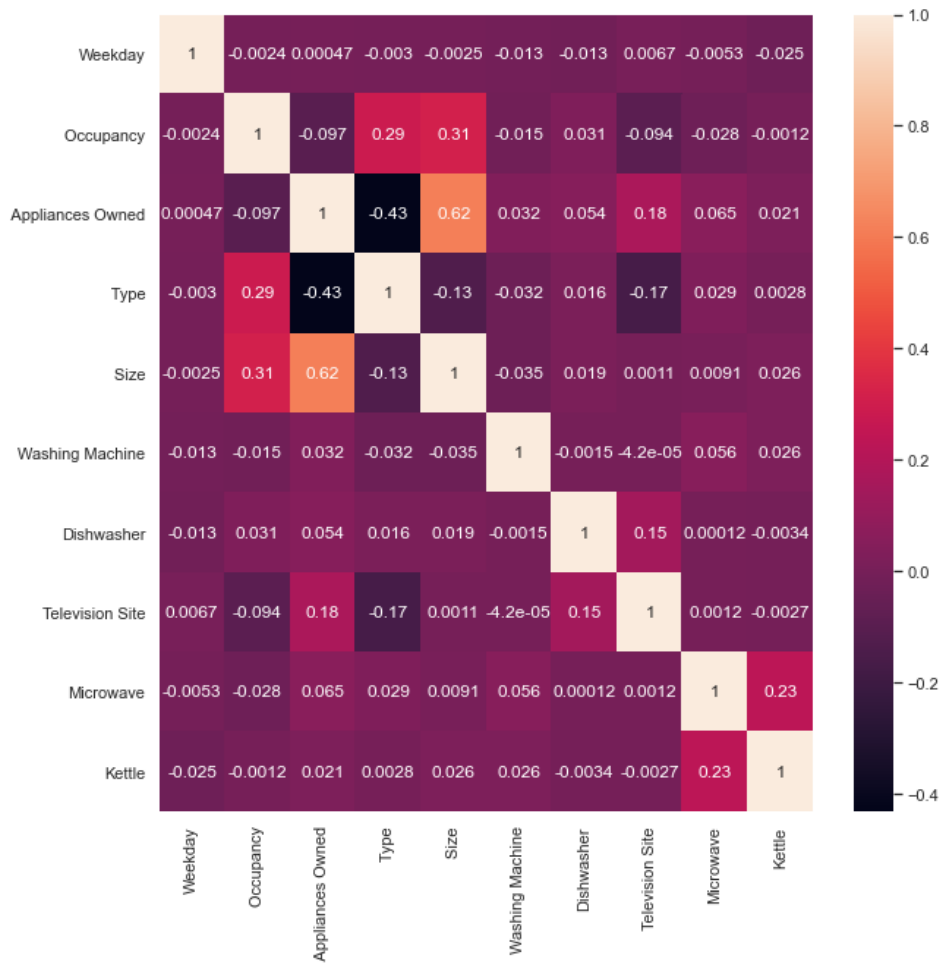


Figure 3.18: Correlation Amongst All the Finalized Features – REFIT

The correlation matrix enables us to look at how one device is being influenced by the other devices and brings light to the fact that device-wise consumption-based prediction is important to understand the overall electricity consumption.

### 3.2.2. Finalizing Data Granularity

Granularity defines the level of detail that is present in a data structure. As we saw in section 3.1.3 and section 3.1.4, the default granularity of REFIT is 3 or 4 seconds and for Sky Electric the granularity is 1s. These values of granularity can capture the slightest details, however, in our current situation we want to capture the on and off events of device. A granularity of this level can be useful in terms of forecasting but, for our classification problem we experiment with the granularity at different levels. Since, we have time series data we shift the granularity up from second to minutes.

The granularities that we can considered are:

#### 1. 1 min

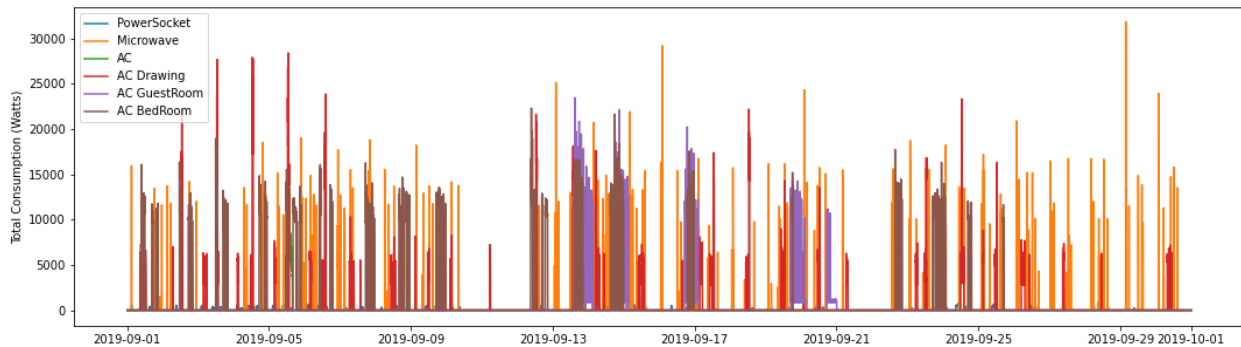


Figure 3.19: Granularity of 1 minute - Sky Electric

#### 2. 5 min

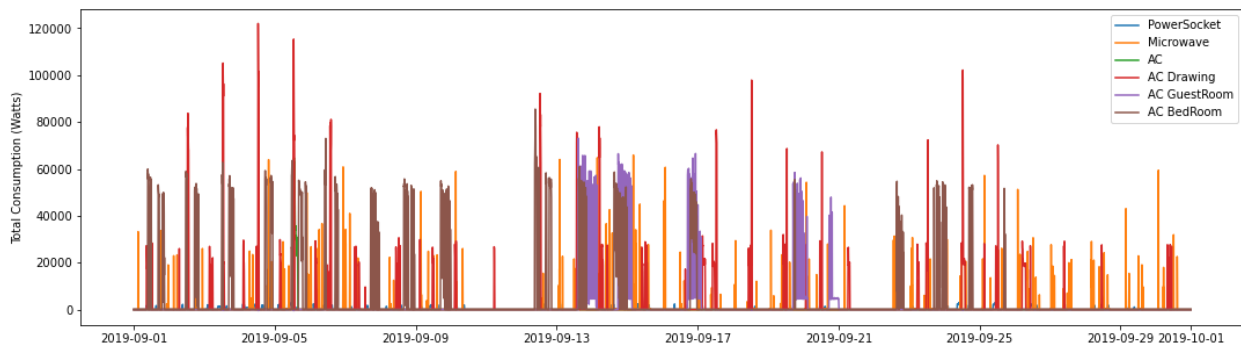


Figure 3.20: Granularity of 5 minutes - Sky Electric

#### 3. 10 min

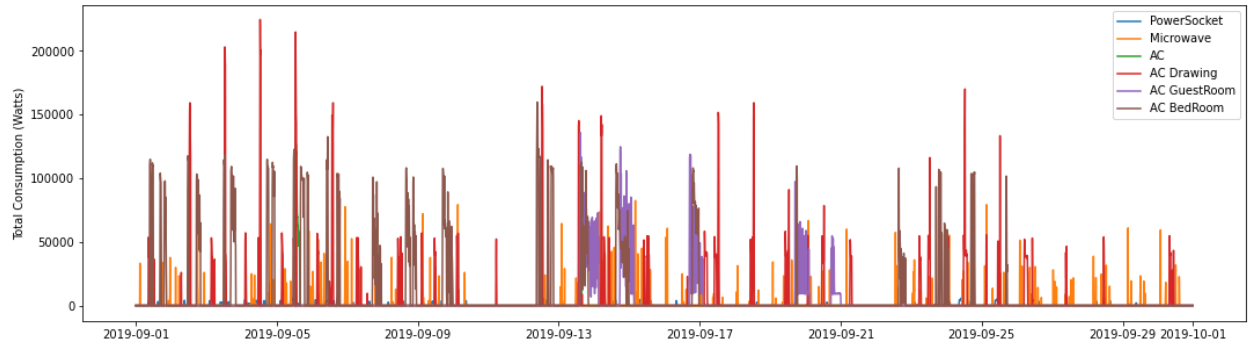


Figure 3.21: Granularity of 10 minutes - Sky Electric

#### 4. 15 min

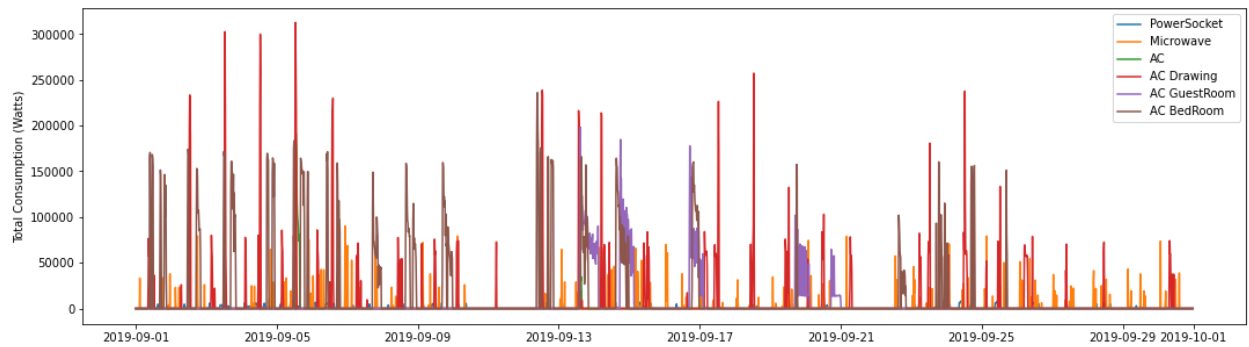


Figure 3.22: Granularity of 15 minutes - Sky Electric

#### 5. 30 min

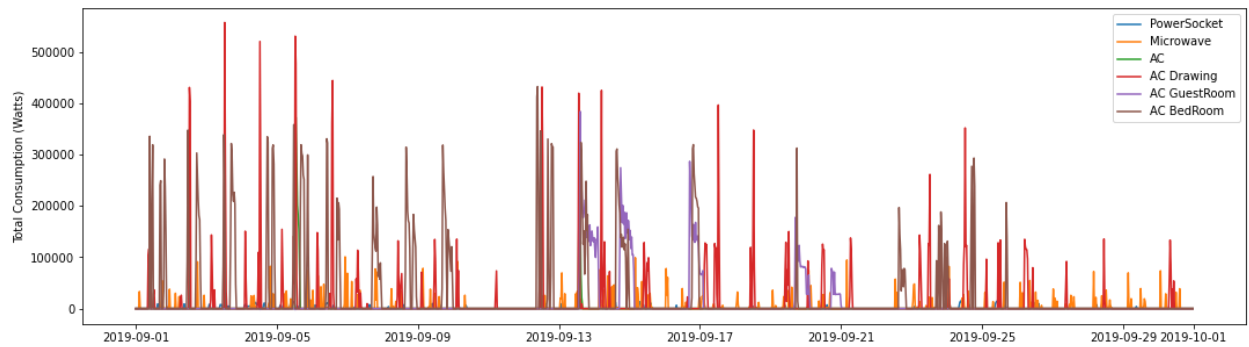


Figure 3.23: Granularity of 15 minutes - Sky Electric

The final granularity that we selected was 10 minutes because, the average duration of consumption of all the devices that we have selected is more than 10 minutes, so, a window of 10 minutes is perfectly able to capture the state of the device. Going above 10 minutes makes the data



very abstract and a lot of details get missed out and going below 10 minutes just adds additional rows which doesn't comply with our research.

### 3.2.3. Altering Data Representation (Converting to Binary Representation)

This is a classification problem, let's say, microwave and AC are turned on at a current point in time then we predict whether the washing machine will be turned on along with them or not. This untold knowledge defines the electricity load profile of a household. Now, we classify the device as either on or off, hence, binary representation.

Each electricity consuming device has 3 stages, which are as follows:

1. Off Stage
2. Standby Stage
3. Consumption Stage

The off stage simply means that the device is turned off, hence the consumption is 0 (off) and it is recorded as such in the dataset. The second stage i.e., the second stage, is a tricky one, even though the devices seem like they are not consuming much power, but it tends to add up quickly and the consumption elevates, hence, we consider as 1 (on). The third stage is the consumption stage, which means that the device is consuming electricity, which is recorded in terms of watts, so another case labeled as 1 (on).

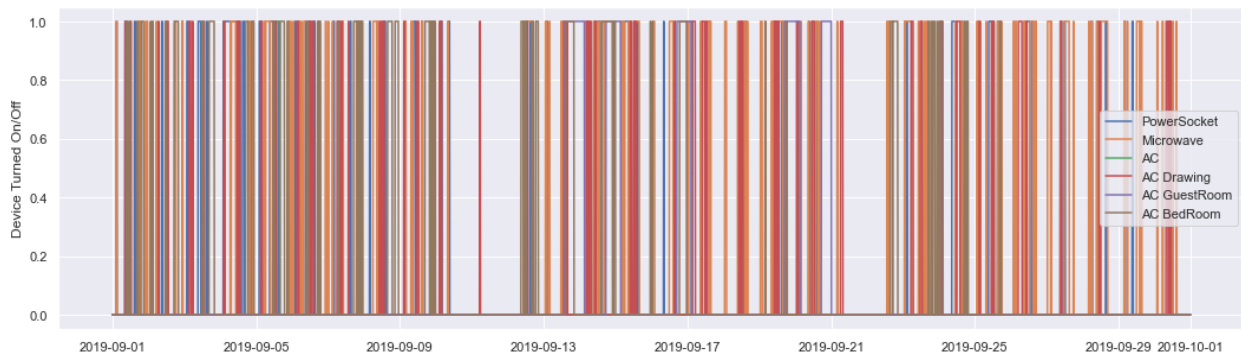


Figure 3.24: Total Device-wise Consumption (Binarized) - Sky Electric

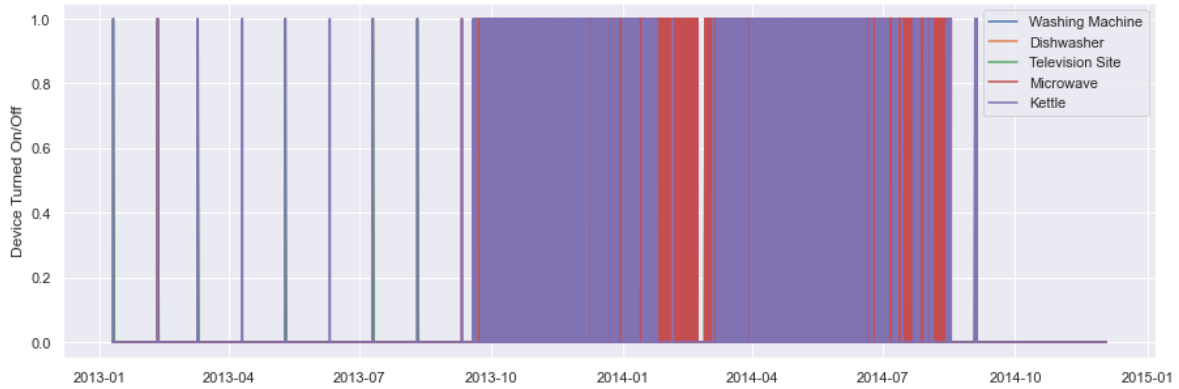


Figure 3.25: Total Device-wise Consumption (Binarized) – REFIT

### 3.3. Post Analysis

Once the required pre-processing has been performed on the dataset, we analyze the data into further details before we start training the model. The post analysis has been carried out by segmenting REFIT into multiple window sizes which represent a specific time window of the day.

#### 3.3.1. Time Window Segmentation

We selected multiple time window to segment the data, each time window has been referenced based on the regular activities that are associated with that time.

##### 1. 12 am to 7 am

For this segment we were hoping to see less usage of all the devices because it's midnight till 7 am. The utilization of TV in the background makes complete sense. The pattern that we see is easily justifiable based on the number of residents in the house. During the night we see the use of dishwasher, it isn't very common but complicit with the residents. The use of kettle after 5 in the morning, is the most reasonable consumption pattern.

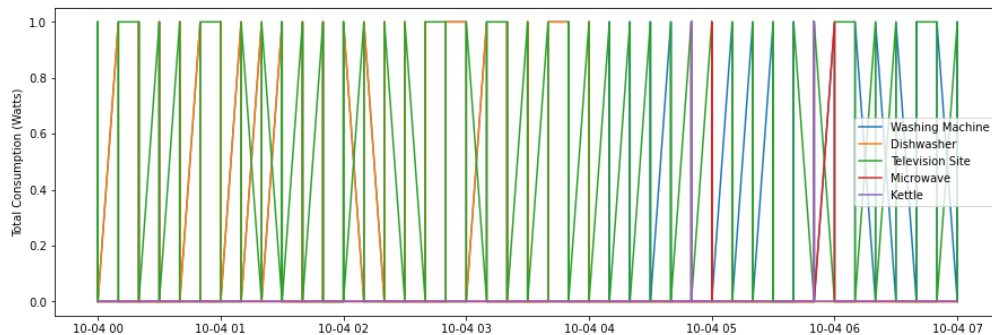


Figure 3.26: REFIT Segmentation Window (12 am - 7 pm)

## 2. 7 am to 10 am

This is a busy time window, usually within this duration the children are leaving for schools etc., the adults leave for offices. So as expected, we see spikes in the usage of kettle, the television has been used minimally, there has been a usage of washing machine, which again coincides with the residents. As expected, we don't see dishwasher being used, this window covers, getting done with what is possible, before leaving, and leaving behind dirty dishes is very common.

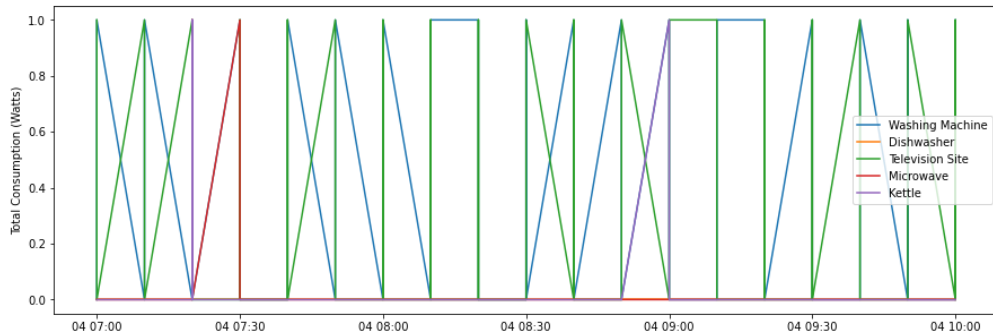


Figure 3.27: REFIT Segmentation Window (7 am - 10 pm)

## 3. 10 am to 3 pm

This time window is considered idle for families with children and working parents, but, if there is a stay-at-home person, then we expect to see usage of almost all the devices. The visualizations show that washing machine, kettle, microwave have been utilized within this window.

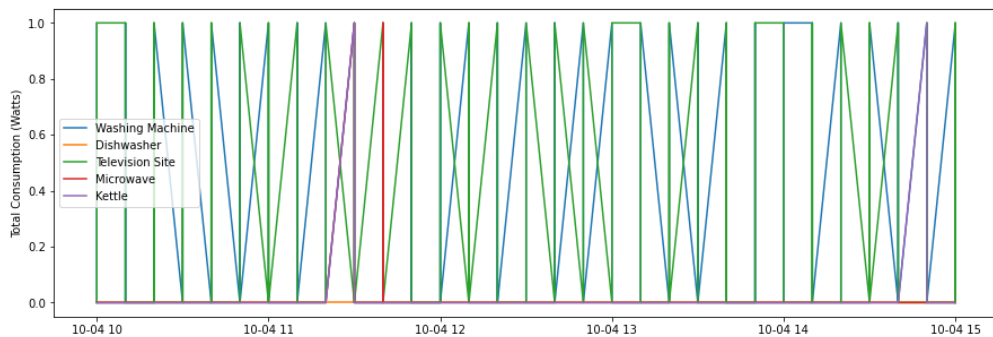


Figure 3.28: REFIT Segmentation Window (10 am - 3 pm)

## 4. 3 pm to 8 pm

This time window is usually the period of the day where most of the residents are at home, we expect to see utilization of all the devices, majorly the television.

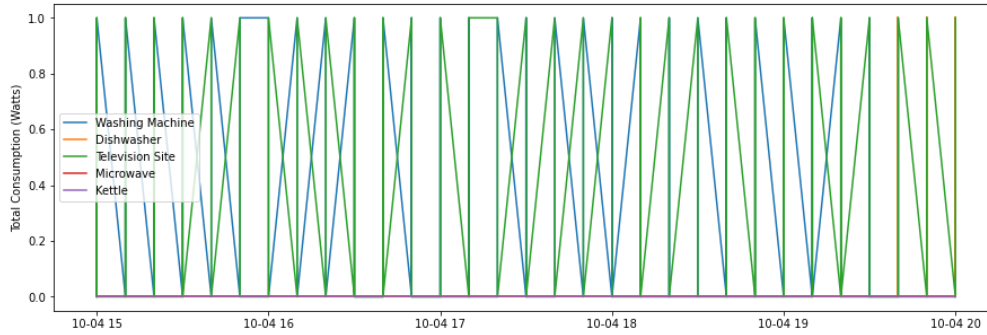


Figure 3.29: REFIT Segmentation Window (3 pm - 8 pm)

## 5. 8pm to 12 am

Another interesting time window where we expect majority of the consumption till 9.30 pm and beyond that, the consumption is usually made by the television devices as part of the leisure.

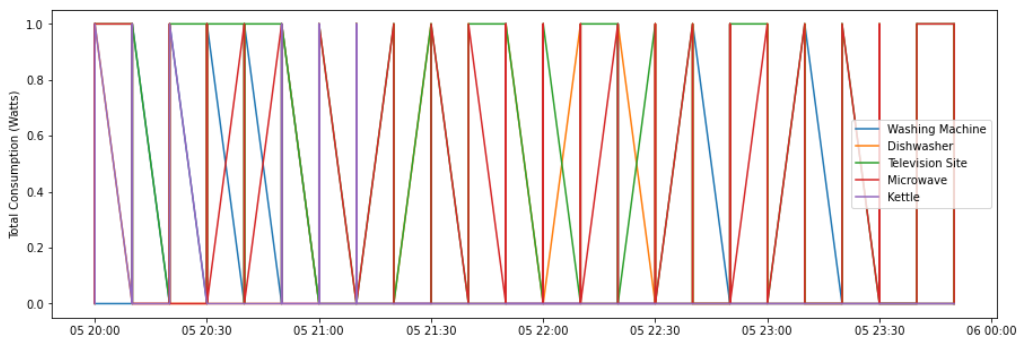


Figure 3.30: REFIT Segmentation Window (8 pm - 12 am)

## 3.4. Machine Learning and Deep Learning

Once the data has been pre-processed and analyzed as per our need, the final step is to train the models that we have selected to generate the electricity load profiles. This is a classification problem; hence we have used 3 different types of algorithms to generate a model for device-wise electricity consumption predictions, that tells us whether a device is on or off, given the status of other devices, the historic features, and the socio-economic features (where they apply). The algorithms used are as follows:

- **Machine Learning**
  - Random Forest Classifier (RF)
  - Hidden Markov Model (HMM)
- **Deep Learning**

- Artificial Neural Network (ANN)

Now, we'll discuss the approaches that we have used, and the way parameters have been defined.

### **3.4.1. Random Forest Classifier**

Random Forest is one of the strongest and most used classifiers that operates using the bagging approach to generate a prediction value. The structure of the model was quite straight forward since we are working on binary prediction and values for only one device are predicted at a time. Hence, the model takes multiple inputs and generates output for a single column.

**Input:** Other device, Weekday, Socio-Economic Parameters

**Output:** Binary Classification of the device being predicted

We had the data split into test and train sets, so that the model can be validated while it is being trained. The test train ratio has been set to 70% from the train set and 30% for the test set respectively. The overall setting of the hyperparameters is as follows:

- Criterion = Gini
- N\_estimators = 100
- max\_depth = 4

For REFIT, the model is run twice, once for data without the socio-economic features and then for data with socio-economic features.

### **3.4.2. Hidden Markov Model**

Hidden Markov Models are popular for training data where sequences exist, they observe the evolutionary behavior in states that is affected by some internal factors but, those factors are not generally observed as something that may impact a change in behavior [46][47]. We use hidden Markov model as a classifier, to classify, that given a certain number of devices that are in either off or on state, what will be the state of the incoming device would it be turned on or off given the effect of the input parameters. HMM has same input and out pattern as random forest classifier[48].

**Input:** Other device, Weekday, Socio-Economic Parameters

**Output:** Binary Classification of the device being predicted

The hyperparameter settings for the model include,

- Normal Distribution
- N\_component = 2

The test train split ratio is also similar as in random forest i.e., 7:3, however, sequence is very important in HMMs hence the division is sequential and not random.

### **3.4.3. Artificial Neural Nets (ANN)**

Though we majorly rely on statistical formulations for generating the electricity load profiles, however, we experiment with a simple deep learning model as well, to compare the results and see which approach provides us with an overall advantage. Since the dataset is not highly complicated and the number of features in both the datasets are limited, ANN felt like the best option [49].

The input/ output paradigm and the train test split of the data is the same as that in random forest classifier. The hyperparameter setting done for the model are as follows:

- Number of dense layers – 3
- Activation Function - ReLU
- Output layer activation function – Sigmoid
- Loss – Binary Cross Entropy
- Optimizer – Adam
- Evaluation Metrics: Accuracy, F1-score

We use sigmoid and binary cross entropy function because our data we require our final output as either 0 or 1.

## **3.5. Evaluation Metrics**

We have selected similar evaluation metrics for all three models, because perform the same task of classifying into binary labels. The selected metric are as follows:

- Area Under the Curve
- F1- Score
  - Precision
  - Recall

These metrics highlight what we want from the model. We target on getting a high value of sensitivity for each device because in a real-time system False Negatives can cost a lot. If a device is off but is predicted as on, this is something that can be coped with but, if it is on and predicted as off then this can disrupt the applications that we plan to solve using the electricity load profiles of the households.

## Chapter 4

### Results

In this section, we will discuss the outcomes of all our experiments. The task at hand was to predict the consumption state of a device given the consumption states of other devices in a household; this is the definition of the household electricity load profiles that we are generating. The results will include the following:

Table 4.1: Summary of Details Covered Under Results

Dataset	Dataset Types	Number of Devices	Algorithm	Evaluation Metrics
REFIT	<ol style="list-style-type: none"><li>1. Prediction based on historic data.</li><li>2. Prediction based on socio-economic features.</li></ol>	5 (Washing Machine, Dishwasher, TV, Microwave, and Kettle)	Random Forest, Hidden Markov Model, Artificial Neural Nets	<ol style="list-style-type: none"><li>1. Area under the curve</li><li>2. F1-score<ol style="list-style-type: none"><li>a. Precision</li><li>b. Recall</li></ol></li></ol>
Sky Electric	Prediction based on historic data.	5 (Microwave, High Voltage Bulb, Power Socket, AC Bedroom, AC Drawing Room)	Random Forest, Hidden Markov Model, Artificial Neural Nets	<ol style="list-style-type: none"><li>1. Area under the curve</li><li>2. F1-score<ol style="list-style-type: none"><li>a. Precision</li><li>b. Recall</li></ol></li></ol>

The results are computed separately for every single device and will be displayed accordingly. We will start step by step and first display all the results for each device monitored under REFIT and followed by that we will do the same for all the monitored devices of Sky Electric dataset.

#### 4.1. Results Compilation for REFIT Dataset

As mentioned in Table 4.1, the total number of devices that we have selected from REFIT are 5. We have predicted each one of those 5 devices by all the three algorithms that we have selected and computed the evaluation metrics, respectively.



### 4.1.1. Without Socio Economic (SE) Features

Without socio-economic features the REFIT is simply just the historic data, based on which the devices have been recorded.

#### 4.1.1.1. Washing Machine

Washing machine was one of the most found device that was monitored under all the 20 REFIT houses. Washing machine is a high voltage energy consumption device and we believe that in many households, it is not run-in combination with other high voltage electricity consumption devices.

##### 4.1.1.1.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.2.

Table 4.2: Random Forest Classifier for Predicting Washing Machine State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.65	
Precision for class Off (0)	0.96	Precision for class On (1)	0.14
Recall for class Off (0)	0.77	Recall for class On (1)	0.54
F1-Score for class Off (0)	0.85	F1-Score for class On (1)	0.23
Total Accuracy		0.75	

##### 4.1.1.1.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.3.

Table 4.3: Hidden Markov Model for Predicting Washing Machine State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.57	
Precision for class Off (0)	0.89	Precision for class On (1)	0.18
Recall for class Off (0)	0.66	Recall for class On (1)	0.48
F1-Score for class Off (0)	0.76	F1-Score for class On (1)	0.26
Total Accuracy		0.64	

#### 4.1.1.1.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.4.

Table 4.4: ANN for Predicting Washing Machine State - No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.58	
Precision for class Off (0)	0.72	Precision for class On (1)	0.13
Recall for class Off (0)	0.51	Recall for class On (1)	0.79
F1-Score for class Off (0)	0.64	F1-Score for class On (1)	0.24
Total Accuracy		0.61	

#### 4.1.1.2. Dishwasher

Dishwasher was also a commonly found device, it was monitored under 80% of the 20 REFIT houses. Dishwasher is a high voltage energy consumption device, and we believe that in many households, it is not run-in combination with other high voltage electricity consumption devices.

##### 4.1.1.2.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.5.

Table 4.5: Random Forest Classifier for Predicting Dishwasher State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.77	
Precision for class Off (0)	0.98	Precision for class On (1)	0.24
Recall for class Off (0)	0.84	Recall for class On (1)	0.72
F1-Score for class Off (0)	0.90	F1-Score for class On (1)	0.36
Total Accuracy		0.83	

#### 4.1.1.2.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.6.

Table 4.6: Hidden Markov Model for Predicting Dishwasher State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.49	
Precision for class Off (0)	0.92	Precision for class On (1)	0.08
Recall for class Off (0)	0.64	Recall for class On (1)	0.35
F1-Score for class Off (0)	0.75	F1-Score for class On (1)	0.13
Total Accuracy		0.62	

#### 4.1.1.2.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned Table 4.7.

Table 4.7: ANN for Predicting Dishwasher State - No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.60	
Precision for class Off (0)	0.78	Precision for class On (1)	0.18
Recall for class Off (0)	0.65	Recall for class On (1)	0.77
F1-Score for class Off (0)	0.72	F1-Score for class On (1)	0.32
Total Accuracy		0.64	

#### 4.1.1.3. Television

Television was also a commonly found device, it was monitored under all the 20 REFIT houses. Television is a low voltage energy consumption device, and usually runs on the standby mode. Hence, we believe that in many households, it is consumed in a combination with other high/low voltage electricity consumption devices.

##### 4.1.1.3.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.8.

Table 4.8: Random Forest Classifier for Predicting Television State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.73	
Precision for class Off (0)	0.83	Precision for class On (1)	0.62
Recall for class Off (0)	0.76	Recall for class On (1)	0.71
F1-Score for class Off (0)	0.79	F1-Score for class On (1)	0.66
Total Accuracy		0.74	

##### 4.1.1.3.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.9.

Table 4.9: Hidden Markov Model for Predicting Television State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.49	
Precision for class Off (0)	0.49	Precision for class On (1)	0.50
Recall for class Off (0)	0.35	Recall for class On (1)	0.64
F1-Score for class Off (0)	0.41	F1-Score for class On (1)	0.56
Total Accuracy		0.50	

#### 4.1.1.3.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.10.

Table 4.10: ANN for Predicting Television State - No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.57	
Precision for class Off (0)	0.67	Precision for class On (1)	0.42
Recall for class Off (0)	0.54	Recall for class On (1)	0.61
F1-Score for class Off (0)	0.72	F1-Score for class On (1)	0.52
Total Accuracy		0.59	

#### 4.1.1.4. Microwave

Microwave was also a commonly found device, it was monitored under 80% of the 20 REFIT houses. Microwave is again, a high voltage energy consumption device. Hence, we believe that in many households, it is not consumed in a combination with other high voltage electricity consumption devices.

##### 4.1.1.4.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.11.

Table 4.11: Random Forest Classifier for Predicting Microwave State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.72	
Precision for class Off (0)	0.84	Precision for class On (1)	0.60
Recall for class Off (0)	0.81	Recall for class On (1)	0.64
F1-Score for class Off (0)	0.83	F1-Score for class On (1)	0.62
Total Accuracy		0.76	

#### 4.1.1.4.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.12.

Table 4.12: Hidden Markov Model for Predicting Microwave State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.50	
Precision for class Off (0)	0.67	Precision for class On (1)	0.34
Recall for class Off (0)	0.64	Recall for class On (1)	0.37
F1-Score for class Off (0)	0.65	F1-Score for class On (1)	0.35
Total Accuracy		0.55	

#### 4.1.1.4.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned Table 4.13.

Table 4.13: ANN for Predicting Microwave State - No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.66	
Precision for class Off (0)	0.68	Precision for class On (1)	0.40
Recall for class Off (0)	0.59	Recall for class On (1)	0.73
F1-Score for class Off (0)	0.63	F1-Score for class On (1)	0.58
Total Accuracy		0.63	

#### 4.1.1.5. Kettle

Kettle was not a commonly found device, it was monitored under 60% of the 20 REFIT houses. But, since it passed 50%, so it made the cut. Kettle, like microwave is again, a high voltage energy consumption device. Hence, we believe that in many households, it is not consumed in a combination with other high voltage electricity consumption devices.

##### 4.1.1.5.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.14.

Table 4.14: Random Forest Classifier for Predicting Kettle State – No SE Features

Without Socio-Economic Parameters			
Area Under the Curve		0.69	
Precision for class Off (0)	0.99	Precision for class On (1)	0.04
Recall for class Off (0)	0.54	Recall for class On (1)	0.84
F1-Score for class Off (0)	0.70	F1-Score for class On (1)	0.07
Total Accuracy		0.54	

##### 4.1.1.5.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.15.

Table 4.15: Hidden Markov Model for Predicting Kettle State – No SE Features

<b>Without Socio-Economic Parameters</b>			
Area Under the Curve		0.50	
Precision for class Off (0)	0.96	Precision for class On (1)	0.04
Recall for class Off (0)	0.71	Recall for class On (1)	0.29
F1-Score for class Off (0)	0.82	F1-Score for class On (1)	0.07
Total Accuracy		0.70	

#### 4.1.1.5.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.16

Table 4.16: ANN for Predicting Kettle State - No SE Features

<b>Without Socio-Economic Parameters</b>			
Area Under the Curve		0.51	
Precision for class Off (0)	0.93	Precision for class On (1)	0.02
Recall for class Off (0)	0.66	Recall for class On (1)	0.75
F1-Score for class Off (0)	0.74	F1-Score for class On (1)	0.05
Total Accuracy		0.68	

### 4.1.2. With Socio Economic (SE) Features

The socio-economic feature, as per our hypothesis help in improving the performance of the models. We saw that they have a correlation with the devices, now, this section, covers the results of the experiments carried out to predict the consumption of all 5 REFIT monitored devices.

#### 4.1.2.1. Washing Machine

Predicting the consumption state of washing machine given the socio-economic features.



#### 4.1.2.1.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.17.

Table 4.17: Random Forest Classifier for Predicting Washing Machine State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.88	
Precision for class Off (0)	0.99	Precision for class On (1)	0.29
Recall for class Off (0)	0.87	Recall for class On (1)	0.90
F1-Score for class Off (0)	0.92	F1-Score for class On (1)	0.44
Total Accuracy		0.87	

#### 4.1.2.1.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.18.

Table 4.18: Hidden Markov Model for Predicting Washing Machine State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.65	
Precision for class Off (0)	0.91	Precision for class On (1)	0.27
Recall for class Off (0)	0.85	Recall for class On (1)	0.71
F1-Score for class Off (0)	0.90	F1-Score for class On (1)	0.56
Total Accuracy		0.72	

#### 4.1.2.1.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.19.

Table 4.19: ANN for Predicting Washing Machine State - SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.88	
Precision for class Off (0)	0.99	Precision for class On (1)	0.25
Recall for class Off (0)	0.83	Recall for class On (1)	0.86
F1-Score for class Off (0)	0.90	F1-Score for class On (1)	0.39
Total Accuracy		0.83	

#### 4.1.2.2. Dishwasher

Predicting the consumption state of dishwasher given the socio-economic features.

##### 4.1.2.2.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.20.

Table 4.20: Random Forest Classifier for Predicting Dishwasher State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.91	
Precision for class Off (0)	0.99	Precision for class On (1)	0.39
Recall for class Off (0)	0.92	Recall for class On (1)	0.92
F1-Score for class Off (0)	0.95	F1-Score for class On (1)	0.55
Total Accuracy		0.92	

##### 4.1.2.2.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.21.

Table 4.21: Hidden Markov Model for Predicting Dishwasher State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.58	
Precision for class Off (0)	0.94	Precision for class On (1)	0.12
Recall for class Off (0)	0.96	Recall for class On (1)	0.45
F1-Score for class Off (0)	0.91	F1-Score for class On (1)	0.35
Total Accuracy		0.75	

#### 4.1.2.2.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.22.

Table 4.22: ANN for Predicting Dishwasher State - SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.89	
Precision for class Off (0)	0.99	Precision for class On (1)	0.31
Recall for class Off (0)	0.87	Recall for class On (1)	0.90
F1-Score for class Off (0)	0.93	F1-Score for class On (1)	0.46
Total Accuracy		0.87	

#### 4.1.2.3. Television

Predicting the consumption state of television given the socio-economic features.

##### 4.1.2.3.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.23.

Table 4.23: Random Forest Classifier for Predicting Television State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.92	
Precision for class Off (0)	0.96	Precision for class On (1)	0.86
Recall for class Off (0)	0.94	Recall for class On (1)	0.91
F1-Score for class Off (0)	0.95	F1-Score for class On (1)	0.89
Total Accuracy		0.93	

#### 4.1.2.3.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.24.

Table 4.24: Hidden Markov Model for Predicting Television State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.61	
Precision for class Off (0)	0.62	Precision for class On (1)	0.65
Recall for class Off (0)	0.55	Recall for class On (1)	0.77
F1-Score for class Off (0)	0.75	F1-Score for class On (1)	0.80
Total Accuracy		0.67	

#### 4.1.2.3.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.25.

Table 4.25: ANN for Predicting Television State - SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.77	
Precision for class Off (0)	0.89	Precision for class On (1)	0.68
Recall for class Off (0)	0.79	Recall for class On (1)	0.82
F1-Score for class Off (0)	0.84	F1-Score for class On (1)	0.74
Total Accuracy		0.80	

#### 4.1.2.4. Microwave

Predicting the consumption state of microwave given the socio-economic features.

##### 4.1.2.4.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.26.

Table 4.26: Random Forest Classifier for Predicting Microwave State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.97	
Precision for class Off (0)	1.0	Precision for class On (1)	0.90
Recall for class Off (0)	0.95	Recall for class On (1)	0.99
F1-Score for class Off (0)	0.97	F1-Score for class On (1)	0.94
Total Accuracy		0.97	

##### 4.1.2.4.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.27.

Table 4.27: Hidden Markov Model for Predicting Microwave State – No SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.56	
Precision for class Off (0)	0.78	Precision for class On (1)	0.42
Recall for class Off (0)	0.65	Recall for class On (1)	0.68
F1-Score for class Off (0)	0.72	F1-Score for class On (1)	0.53
Total Accuracy		0.61	

#### 4.1.2.4.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.28.

Table 4.28: ANN for Predicting Microwave State - SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.91	
Precision for class Off (0)	0.93	Precision for class On (1)	0.56
Recall for class Off (0)	0.71	Recall for class On (1)	0.87
F1-Score for class Off (0)	0.80	F1-Score for class On (1)	0.68
Total Accuracy		0.76	

#### 4.1.2.5. Kettle

Predicting the consumption state of kettle given the socio-economic features.

##### 4.1.2.5.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.29.

Table 4.29: Random Forest Classifier for Predicting Kettle State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.78	
Precision for class Off (0)	1.0	Precision for class On (1)	0.03
Recall for class Off (0)	0.74	Recall for class On (1)	0.83
F1-Score for class Off (0)	0.85	F1-Score for class On (1)	0.05
Total Accuracy		0.74	

#### 4.1.2.5.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.30.

Table 4.30: Hidden Markov Model for Predicting Kettle State – SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.61	
Precision for class Off (0)	0.93	Precision for class On (1)	0.07
Recall for class Off (0)	0.75	Recall for class On (1)	0.35
F1-Score for class Off (0)	0.83	F1-Score for class On (1)	0.10
Total Accuracy		0.68	

#### 4.1.2.5.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.31.

Table 4.31: ANN for Predicting Kettle State - SE Features

With Socio-Economic Parameters			
Area Under the Curve		0.70	
Precision for class Off (0)	1.00	Precision for class On (1)	0.04
Recall for class Off (0)	0.72	Recall for class On (1)	0.83
F1-Score for class Off (0)	0.83	F1-Score for class On (1)	0.07
Total Accuracy		0.72	

This was a compilation of the results that we have generated after experimenting on REFIT. Based on the results of all the devices, using dataset without the socio-economic features and using the dataset with the socio-economic features it is very clear that the performance has improved with the addition of the socio-economic features.

## 4.2. Results Compilation for Sky Electric Dataset

As discussed previously, sky electric is a small privately owned Pakistan-based dataset. We have used it to analyze how well can we predict the device-wise consumption in Pakistani Residential setup. The dataset doesn't provide the socio-economic features hence, the prediction results are only based on the historic data. We have 5 total devices, following are the results that we were able to generate using sky electric dataset.

### 4.2.1. Microwave

Microwave is a very common device which is found in majority of the average Pakistani households. It is a high voltage power consumption device that is a part of the everyday life. Our intuition is that in a regular Pakistani residential settlement, the use of microwave with other high voltage devices is not common. We were able to predict microwave's consumption status using the historical data through all three of the models that we have created.

#### 4.2.1.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.32.



Table 4.32: Random Forest Classifier for Predicting Microwave State

Random Forest Classifier			
Area Under the Curve		0.53	
Precision for class Off (0)	0.21	Precision for class On (1)	0.83
Recall for class Off (0)	0.54	Recall for class On (1)	0.53
F1-Score for class Off (0)	0.30	F1-Score for class On (1)	0.64
Total Accuracy		0.53	

#### 4.2.1.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.33.

Table 4.33: Hidden Markov Model for Predicting Microwave State

Hidden Markov Model			
Area Under the Curve		0.48	
Precision for class Off (0)	0.19	Precision for class On (1)	0.84
Recall for class Off (0)	0.55	Recall for class On (1)	0.49
F1-Score for class Off (0)	0.29	F1-Score for class On (1)	0.62
Total Accuracy		0.50	

#### 4.2.1.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.34.

Table 4.34: ANN for Predicting Microwave State

Artificial Neural Net			
Area Under the Curve		0.51	
Precision for class Off (0)	0.25	Precision for class On (1)	0.77
Recall for class Off (0)	0.55	Recall for class On (1)	0.45
F1-Score for class Off (0)	0.33	F1-Score for class On (1)	0.63
Total Accuracy		0.54	

## 4.2.2. Power Socket

The power sockets do not consume any electricity unless a device is not connected to them. Sockets are one of the most common sites of electricity consumption. Currently, it is hard for us to predict the device that is connected to the socket, however, we through our current approach, we can tell whether a socket is being used or not i.e., it's turned on or off.

### 4.2.2.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.35.

Table 4.35: Random Forest Classifier for Predicting Power Socket State

Random Forest Classifier			
Area Under the Curve		0.64	
Precision for class Off (0)	0.93	Precision for class On (1)	0.19
Recall for class Off (0)	0.68	Recall for class On (1)	0.61
F1-Score for class Off (0)	0.78	F1-Score for class On (1)	0.29
Total Accuracy		0.67	

#### 4.2.2.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.36.

Table 4.36: Hidden Markov Model for Predicting Power Socket State

Hidden Markov Model			
Area Under the Curve		0.61	
Precision for class Off (0)	0.89	Precision for class On (1)	0.21
Recall for class Off (0)	0.68	Recall for class On (1)	0.57
F1-Score for class Off (0)	0.75	F1-Score for class On (1)	0.32
Total Accuracy		0.63	

#### 4.2.2.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.37.

Table 4.37: ANN for Predicting Power Socket State

Artificial Neural Net			
Area Under the Curve		0.62	
Precision for class Off (0)	0.88	Precision for class On (1)	0.20
Recall for class Off (0)	0.65	Recall for class On (1)	0.58
F1-Score for class Off (0)	0.77	F1-Score for class On (1)	0.27
Total Accuracy		0.65	

### 4.2.3. High Voltage Bulb

The lightning devices that have a very low and constant consumption are termed under cold appliances in this research. Though this bulb comes under lightning devices, but it has a high electricity consumption power and is not being constantly used by the household hence, we consider it for the prediction purposes.

#### 4.2.3.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.38.

Table 4.38: Random Forest Classifier for Predicting HV Bulb State

Random Forest Classifier			
Area Under the Curve		0.59	
Precision for class Off (0)	0.88	Precision for class On (1)	0.23
Recall for class Off (0)	0.54	Recall for class On (1)	0.65
F1-Score for class Off (0)	0.67	F1-Score for class On (1)	0.34
Total Accuracy		0.56	

#### 4.2.3.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.39.

Table 4.39: Hidden Markov Model for Predicting HV Bulb State

Hidden Markov Model			
Area Under the Curve		0.55	
Precision for class Off (0)	0.89	Precision for class On (1)	0.22
Recall for class Off (0)	0.34	Recall for class On (1)	0.62
F1-Score for class Off (0)	0.59	F1-Score for class On (1)	0.33
Total Accuracy		0.51	

#### 4.2.3.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.40.

Table 4.40: ANN for Predicting HV Bulb State

Artificial Neural Net			
Area Under the Curve		0.58	
Precision for class Off (0)	0.82	Precision for class On (1)	0.18
Recall for class Off (0)	0.55	Recall for class On (1)	0.59
F1-Score for class Off (0)	0.60	F1-Score for class On (1)	0.21
Total Accuracy		0.52	

#### 4.2.4. AC (Drawing Room)

The air conditioners are the most found devices in Pakistan with a very high usage during the summer months. ACs are high voltage devices and in an average Pakistani household it's use with other high voltage devices is very low.

##### 4.2.4.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.41.

Table 4.41: Random Forest Classifier for Predicting AC (Drawing Room) State

Random Forest Classifier			
Area Under the Curve		0.56	
Precision for class Off (0)	0.94	Precision for class On (1)	0.09
Recall for class Off (0)	0.59	Recall for class On (1)	0.54
F1-Score for class Off (0)	0.73	F1-Score for class On (1)	0.16
Total Accuracy		0.59	

#### 4.2.4.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.42.

Table 4.42: Hidden Markov Model for Predicting AC (Drawing Room) State

Hidden Markov Model			
Area Under the Curve		0.52	
Precision for class Off (0)	0.90	Precision for class On (1)	0.12
Recall for class Off (0)	0.44	Recall for class On (1)	0.49
F1-Score for class Off (0)	0.68	F1-Score for class On (1)	0.13
Total Accuracy		0.55	

#### 4.2.4.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.43.

Table 4.43: ANN for Predicting AC (Drawing Room) State

Artificial Neural Net			
Area Under the Curve		0.53	
Precision for class Off (0)	0.91	Precision for class On (1)	0.03
Recall for class Off (0)	0.54	Recall for class On (1)	0.49
F1-Score for class Off (0)	0.71	F1-Score for class On (1)	0.12
Total Accuracy		0.57	

## 4.2.5. AC (Bedroom)

### 4.2.5.1. Random Forest Classifier

With Random Forest Classifier, we were able to generate the results as mentioned in Table 4.44.

Table 4.44: Random Forest Classifier for Predicting AC (Bedroom) State

Random Forest Classifier			
Area Under the Curve		0.65	
Precision for class Off (0)	0.96	Precision for class On (1)	0.16
Recall for class Off (0)	0.50	Recall for class On (1)	0.81
F1-Score for class Off (0)	0.65	F1-Score for class On (1)	0.27
Total Accuracy		0.53	

### 4.2.5.2. Hidden Markov Model

With Hidden Markov Model as a classifier, we were able to generate the results as mentioned in Table 4.45.



Table 4.45: Hidden Markov Model for Predicting AC (Bedroom) State

Hidden Markov Model			
Area Under the Curve		0.63	
Precision for class Off (0)	0.92	Precision for class On (1)	0.13
Recall for class Off (0)	0.41	Recall for class On (1)	0.78
F1-Score for class Off (0)	0.52	F1-Score for class On (1)	0.25
Total Accuracy		0.49	

### 4.2.5.3. Artificial Neural Nets

With Artificial Neural Nets as a classifier, we were able to generate the results as mentioned in Table 4.46.

Table 4.46: ANN for Predicting AC (Bedroom) State

Artificial Neural Net			
Area Under the Curve		0.63	
Precision for class Off (0)	0.88	Precision for class On (1)	0.15
Recall for class Off (0)	0.45	Recall for class On (1)	0.79
F1-Score for class Off (0)	0.55	F1-Score for class On (1)	0.26
Total Accuracy		0.51	

## 4.3. Comparison of Approaches

Now that the results have been formulated for all the devices based on the 3 approaches that we have designed, we compare the results.

The initial comparison in Table 4.47 highlights which dataset type (with socio-economic features or without socio economic features) was able to provide the best results.

Table 4.47: Comparison of all REFIT Devices based on the Results from all Three Approaches and Type of the Dataset (with SE or Without SE). Results Shown in %

<b>Device</b>	<b>Classifier</b>	<b>Without Socio- Economic Features</b>	<b>With Socio- Economic Features</b>
<b>Washing Machine</b>	Random Forest	75%	<b>87%</b>
	Hidden Markov Model	64%	72%
	Artificial Neural Net	61%	83%
<b>Dishwasher</b>	Random Forest	83%	<b>92%</b>
	Hidden Markov Model	62%	75%
	Artificial Neural Net	64%	87%
<b>Television</b>	Random Forest	74%	<b>93%</b>
	Hidden Markov Model	50%	67%
	Artificial Neural Net	59%	80%
<b>Microwave</b>	Random Forest	76%	<b>97%</b>
	Hidden Markov Model	55%	61%
	Artificial Neural Net	63%	76%
<b>Kettle</b>	Random Forest	54%	<b>74%</b>
	Hidden Markov Model	70%	68%
	Artificial Neural Net	68%	72%

The next comparison is to show which approach was able to outperform the other approaches, given the data only has historical data and no socio-economic features.

Table 4.48: A Comparison of all the Non-SE data (REFIT and Sky Electric) to compare the three approaches.  
Results Shown in %

<b>Device</b>	<b>Random Forest Classifier</b>	<b>Hidden Markov Model</b>	<b>Artificial Neural Network</b>
<b>Washing Machine</b>	<b>75%</b>	64%	61%
<b>Dishwasher</b>	<b>83%</b>	62%	64%
<b>Television</b>	<b>74%</b>	50%	59%
<b>Microwave (REFIT)</b>	<b>76%</b>	55%	63%
<b>Kettle</b>	54%	70%	<b>68%</b>
<b>Microwave (Sky Elec.)</b>	53%	50%	<b>54%</b>
<b>Power Socket</b>	<b>67%</b>	63%	65%
<b>High Voltage Bulb</b>	<b>56%</b>	51%	52%
<b>AC (Drawing Room)</b>	<b>59%</b>	55%	57%
<b>AC (Bedroom)</b>	<b>53%</b>	49%	51%

## Chapter 5

### Discussions

This chapter majorly focusses on the findings that we have discovered over the course of this research. In addition, we point out the current limitations of our work.

#### 5.1. Summary of the Findings

The target of this research was to be able to generate electricity load profiles that captures the device-wise consumption pattern of residential households.

Now, we discuss the results that we were able to generate. We decided on using two machine learning (statistical) models and 1 deep learning model. The initial comparison was to prove our hypothesis that the socio-economic features enhance the overall prediction accuracy of the devices. A device's consumption status is highly reliant on the residents that are using the device. For example, if there is 1 person living in an apartment who goes to work at 9 am and returns at 5pm, then there will be no one left to consume any devices except the cold appliances. The relationship between the usage and the demographics is very clear. We only had **socio-economic features** provided with the REFIT, so we computed the prediction scores for all the devices, once without the socio-economic features and once with them. We were able to achieve maximum prediction accuracy for the microwave which was **97%** when predicted on a dataset that had socio-economic features merged. On the other hand, **without socioeconomic features** the prediction accuracy was **76%**. So, we see a clear jump of 26% in the overall accuracy of microwave using random forest classifier. Similarly, through HMM and ANN the accuracy of the microwave was higher with the SE features. The results followed in a similar pattern for Washing Machine, Television and Dishwasher and Kettle. This completely complies with the theory that the consumption pattern of the devices is associated with the household's demography. Washing machine and dishwasher are the two devices that are highly dependent on the number of residents in a household. Higher the number of residents, the more frequent would be the usage of these two devices. Television, Kettle, and Microwave, go both ways. 1 person could use them just as much as a family would, they are completely dependent on user's lifestyle. The comparative analysis has been given in

Following this, the next comparison was to choose which model is the best in terms of performance. We performed this comparative analysis using all the devices from both REFIT and

Sky Electric data and made predictions without considering the socio-economic features. Random Forest classifier was coined as the best classifier for predicting device-wise electricity consumption, as it outperformed both Hidden Markov Model and Artificial Neural Network in 8 out of 10 total devices. The second in line was the ANN, however, the HMM could not make its mark given the current condition of the data.

The reason we believe why the ANN was not able to perform as well as the Random Forest is because the data is not large enough to allow the model to train accurately.

This comparative analysis has been shown in

## **5.2. Comparison with Existing Work**

Through Chapter 2, we were able to review the existing literature in the domain of generating electricity load profiles. Majority of the work currently being done, targets the infrastructure of developed countries, where the consumption of electricity is being monitored to some extent. However, in Pakistan, we lack the basic monitoring. Our hypothesis has shown how device-wise prediction of electricity consumption provides us with better insights on understanding the consumption pattern of a household which has previously been done using minimalistic devices like lightning device [39]. Also, we have shown how the introduction of socio-economic parameters can further improve the load profiles by comparing the results against only historic data. The results show that our hypothesis is a success and device-wise consumption can be predicted more accurately with the introduction of the socio-economic parameters.

## **5.3. Limitations**

Our electricity load profiles are dependent upon the existence of device-wise data, however, majority of the publicly available datasets that target problem solving in the domain of residential electricity consumption, don't provide data from many household appliances. In our current situation, we selected REFIT as the major research dataset because, it covered all the 3 things that we require in a dataset i.e., historic data, device-wise monitored data, and socio-economic features of the households. The data was recorded over a period of 2 years, so, it was capturing the seasonality. However, the recorded devices in almost all the 20 houses of the dataset were different. To finalize the data, we had to select the devices that were common in all the houses, the number was small and hence we also had to select devices that were found in majority of the houses. This

meant, that for all the houses where that specific device was not present the value would be recorded as 0 and the device would be considered off. This cause extreme class imbalance, specifically in devices that don't have a prolonged consumption time.

Without Socio-Economic Parameters				With Socio-Economic Parameters			
Area Under the Curve		0.69		Area Under the Curve		0.78	
Precision for class Off (0)	0.99	Precision for class On (1)	0.04	Precision for class Off (0)	1.0	Precision for class On (1)	0.03
Recall for class Off (0)	0.54	Recall for class On (1)	0.84	Recall for class Off (0)	0.74	Recall for class On (1)	0.83
F1-Score for class Off (0)	0.70	F1-Score for class On (1)	0.07	F1-Score for class Off (0)	0.85	F1-Score for class On (1)	0.05
Total Accuracy		0.54		Total Accuracy		0.74	

Figure 5.1: Fall in the Precision of Kettle Due to Class Imbalance

Through Figure 5.1, we can see that the results of the kettle, in terms of precision of class 1 are extremely affected because of the huge imbalance.

Another limitation is that most of the household residents consider sharing the socio-economic features as a breach to their privacy and hence avoid sharing information [16][50]. This makes us lose a lot of houses that could otherwise be sources of data.

## Chapter 6

### Conclusions

This chapter concludes all the research work done under this thesis, by describing the contributions and indicating the direction for future work in the domain of generating electricity load profiles based on device-wise consumption pattern. First, we will talk about the core contributions of our research and then highlight the possible dimensions of the future work.

#### 6.1. Summary of Research Contributions

In terms of electricity consumption, knowing the behavior of every individual device is extremely important because it cultivates the entire electricity consumption pattern. If we can correctly identify when devices are being utilized in a household and under which circumstances, then we can establish limits to the amount of electricity that is being supplied to the household. In majority of the current literature, the load profiles of residential households have not been generated on such a small scale. Though the entire idea of the research is not new, however, creating electricity load profiles by considering device-wise consumption and the socio-economic features in one single model, is a novel approach.

To understand the relation between socio-economic features and the electricity consumption data we analyzed the data by segmenting it into smaller time windows, to get a better understanding. To further emphasize that the socio-economic features add an improvement to the overall device-wise predictions, we have computed predictions on two different machine learning models i.e., Random Forest Classifier, Hidden Markov Model and, one deep learning model i.e., Artificial Neural Network for two types of datasets. One in which REFIT is associated with socio-economic features and one in which it is not associated. Then to finalize the which model performs best, we have predicted device state based only on historical data. This has been done for both REFIT and Sky Electric data. We have been able to achieve outstanding accuracy, e.g., on microwave, random forest classifier provided an accuracy of **97%** with the socio-economic features and **76%** without the socio-economic features. Similarly, dishwasher was predicted up to an accuracy of **92%** with SE features and **84%** without the SE features.

## **6.2. Future Work**

The possible dimensions that we foresee as the future work of this research, are as follows:

- The datasets that we had access to for this research were limited, in terms of size and the devices provided. Also, the datasets lack consistency in the devices that have been monitored from each household. In addition to these, we require more datasets that have the socio-economic features provided. The creation of such large datasets is important to take this research a step further.
- We have currently worked with only one deep learning model; however, we can try training with multiple, more complex deep learning models to see how they perform.
- The research can be taken one step further, and the generated load profiles can be used to forecast early load shedding.



## References

- [1] L. E. Doman *et al.*, “World energy demand and economic outlook Macroeconomic assumptions Liquid fuels Natural gas Transportation sector Energy-related carbon dioxide emissions The following also contributed to the production of the IEO2013 report,” 2040. [Online]. Available: [www.eia.gov](http://www.eia.gov)
- [2] M. T. Jaiyesimi, T. S. Osinubi, and L. Amaghionyeodiwe, “Energy Consumption and GGP in the OECD Countries: A Causality Analysis,” *Review of Economic and Business Studies*, vol. 10, no. 1, pp. 55–74, Jun. 2017, doi: 10.1515/rebs-2017-0048.
- [3] L. Sun *et al.*, “Optimisation model for power system restoration with support from electric vehicles employing battery swapping,” *IET Generation, Transmission and Distribution*, vol. 10, no. 3, pp. 771–779, Feb. 2016, doi: 10.1049/iet-gtd.2015.0441.
- [4] N. Oconnell, P. Pinson, H. Madsen, and M. Omalley, “Benefits and challenges of electrical demand response: A critical review,” *Renewable and Sustainable Energy Reviews*, vol. 39. Elsevier Ltd, pp. 686–699, 2014. doi: 10.1016/j.rser.2014.07.098.
- [5] M. Shahbaz, A. R. Chaudhary, and I. Ozturk, “Munich Personal RePEc Archive Does urbanization cause increasing energy demand in Pakistan? Empirical evidence from STIRPAT model Does urbanization cause increasing energy demand in Pakistan? Empirical evidence from STIRPAT model,” 2016.
- [6] “The Relationship between Electricity Consumption, Oil Prices, and Economic Growth in Indonesia.” [Online]. Available: <http://journal2.uad.ac.id/index.php/jampe/index>
- [7] K. Khan, A. Shah, and J. Khan, “Electricity Consumption Patterns: Comparative Evidence from Pakistan’s Public and Private Sectors.” [Online]. Available: <http://tribune.com.pk/story/154420/countrywide-energy-shortage-as-pepco-increases->
- [8] H. Allcott *et al.*, “NBER WORKING PAPER SERIES HOW DO ELECTRICITY SHORTAGES AFFECT INDUSTRY? EVIDENCE FROM INDIA Previously circulated as ‘How Do Electricity Shortages Affect Productivity? Evidence from India.’ We thank How Do Electricity Shortages Affect Industry? Evidence from India How Do Electricity

- Shortages Affect Industry? Evidence from India,” 2014. [Online]. Available: <http://www.nber.org/papers/w19977>
- [9] T. Baskaran, B. Min, and Y. Uppal, “Election cycles and electricity provision: Evidence from a quasi-experiment with Indian special elections,” *Journal of Public Economics*, vol. 126, pp. 64–73, Jun. 2015, doi: 10.1016/j.jpubeco.2015.03.011.
- [10] D. Fischer, A. Härtl, and B. Wille-Hausmann, “Model for electric load profiles with high time resolution for German households,” *Energy and Buildings*, vol. 92, pp. 170–179, Apr. 2015, doi: 10.1016/j.enbuild.2015.01.058.
- [11] C. Wolfram, O. Shelef, and P. Gertler, “How will energy demand develop in the developing world?,” in *Journal of Economic Perspectives*, Dec. 2012, vol. 26, no. 1, pp. 119–138. doi: 10.1257/jep.26.1.119.
- [12] I. Mezghani and H. ben Haddad, “Energy consumption and economic growth: An empirical study of the electricity consumption in Saudi Arabia,” *Renewable and Sustainable Energy Reviews*, vol. 75. Elsevier Ltd, pp. 145–156, 2017. doi: 10.1016/j.rser.2016.10.058.
- [13] Y. Wang, H. Lin, Y. Liu, Q. Sun, and R. Wennersten, “Management of household electricity consumption under price-based demand response scheme,” *Journal of Cleaner Production*, vol. 204, pp. 926–938, Dec. 2018, doi: 10.1016/j.jclepro.2018.09.019.
- [14] P. Escobar, E. Martínez, J. C. Saenz-Díez, E. Jiménez, and J. Blanco, “Modeling and analysis of the electricity consumption profile of the residential sector in Spain,” *Energy and Buildings*, vol. 207, Jan. 2020, doi: 10.1016/j.enbuild.2019.109629.
- [15] Y. Ding, Q. Wang, Z. Wang, S. Han, and N. Zhu, “An occupancy-based model for building electricity consumption prediction: A case study of three campus buildings in Tianjin,” *Energy and Buildings*, vol. 202, Nov. 2019, doi: 10.1016/j.enbuild.2019.109412.
- [16] B. Völker, A. Reinhardt, A. Faustine, and L. Pereira, “Watt’s up at home? Smart meter data analytics from a consumer-centric perspective,” *Energies*, vol. 14, no. 3. MDPI AG, Feb. 01, 2021. doi: 10.3390/en14030719.

- [17] M. N. Fekri, H. Patel, K. Grolinger, and V. Sharma, “Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network,” *Applied Energy*, vol. 282, Jan. 2021, doi: 10.1016/j.apenergy.2020.116177.
- [18] M. A. Shaukat, H. R. Shaukat, Z. Qadir, H. S. Munawar, A. Z. Kouzani, and M. A. P. Mahmud, “Cluster analysis and model comparison using smart meter data,” *Sensors*, vol. 21, no. 9, May 2021, doi: 10.3390/s21093157.
- [19] K. R. Abbasi, J. Abbas, S. Mahmood, and M. Tufail, “Revisiting electricity consumption, price, and real GDP: A modified sectoral level analysis from Pakistan,” *Energy Policy*, vol. 149, Feb. 2021, doi: 10.1016/j.enpol.2020.112087.
- [20] M. Fahim and A. Sillitti, “Analyzing load profiles of energy consumption to infer household characteristics using smart meters,” *Energies (Basel)*, vol. 12, no. 5, Feb. 2019, doi: 10.3390/en12050773.
- [21] S. Singh and A. Yassine, “Big data mining of energy time series for behavioral analytics and energy consumption forecasting,” *Energies (Basel)*, vol. 11, no. 2, Feb. 2018, doi: 10.3390/en11020452.
- [22] A. Nadeem and N. Arshad, “PRECON: Pakistan residential electricity consumption dataset,” in *e-Energy 2019 - Proceedings of the 10th ACM International Conference on Future Energy Systems*, Jun. 2019, pp. 52–57. doi: 10.1145/3307772.3328317.
- [23] L. Nikdel, A. E. S. Schay, D. Hou, and S. E. Powers, “Data-driven Occupancy Profiles for Apartment-style Student Housing,” *Energy and Buildings*, vol. 246, Sep. 2021, doi: 10.1016/j.enbuild.2021.111070.
- [24] H. Li, Z. Wang, T. Hong, A. Parker, and M. Neukomm, “Characterizing patterns and variability of building electric load profiles in time and frequency domains,” *Applied Energy*, vol. 291, Jun. 2021, doi: 10.1016/j.apenergy.2021.116721.
- [25] S. Vojtovic, A. Stundziene, and R. Kontautiene, “The impact of socio-economic indicators on sustainable consumption of domestic electricity in Lithuania,” *Sustainability (Switzerland)*, vol. 10, no. 2, Jan. 2018, doi: 10.3390/su10020162.

- [26] B. Jeong, J. Kim, and R. de Dear, "Creating household occupancy and energy behavioural profiles using national time use survey data," *Energy and Buildings*, vol. 252, Dec. 2021, doi: 10.1016/j.enbuild.2021.111440.
- [27] T. Ahmad *et al.*, "Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment," *Energy*, vol. 158, pp. 17–32, Sep. 2018, doi: 10.1016/j.energy.2018.05.169.
- [28] I. Mahmood, Quair-tul-ain, H. A. Nasir, F. Javed, and J. A. Aguado, "A hierarchical multi-resolution agent-based modeling and simulation framework for household electricity demand profile," *Simulation*, vol. 96, no. 8, pp. 655–678, Aug. 2020, doi: 10.1177/0037549720923401.
- [29] A. Satre-Meloy, M. Diakonova, and P. Grünewald, "Cluster analysis and prediction of residential peak demand profiles using occupant activity data," *Applied Energy*, vol. 260, Feb. 2020, doi: 10.1016/j.apenergy.2019.114246.
- [30] L. Sun, K. Zhou, and S. Yang, "An ensemble clustering based framework for household load profiling and driven factors identification," *Sustainable Cities and Society*, vol. 53, Feb. 2020, doi: 10.1016/j.scs.2019.101958.
- [31] G. Trotta, "An empirical analysis of domestic electricity load profiles: Who consumes how much and when?," *Applied Energy*, vol. 275, Oct. 2020, doi: 10.1016/j.apenergy.2020.115399.
- [32] N. Elamin and M. Fukushige, "Modeling and forecasting hourly electricity demand by SARIMAX with interactions," *Energy*, vol. 165, pp. 257–268, Dec. 2018, doi: 10.1016/j.energy.2018.09.157.
- [33] X. Liu, Y. Ding, H. Tang, and F. Xiao, "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data," *Energy and Buildings*, vol. 231, Jan. 2021, doi: 10.1016/j.enbuild.2020.110601.

- [34] N. M. M. Bendaoud, N. Farah, and S. ben Ahmed, "Applying load profiles propagation to machine learning based electrical energy forecasting," *Electric Power Systems Research*, vol. 203, Feb. 2022, doi: 10.1016/j.epsr.2021.107635.
- [35] J. S. Chou and D. S. Tran, "Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders," *Energy*, vol. 165, pp. 709–726, Dec. 2018, doi: 10.1016/j.energy.2018.09.144.
- [36] N. Mohan, K. P. Soman, and S. Sachin Kumar, "A data-driven strategy for short-term electric load forecasting using dynamic mode decomposition model," *Applied Energy*, vol. 232, pp. 229–244, Dec. 2018, doi: 10.1016/j.apenergy.2018.09.190.
- [37] C. Behm, L. Nolting, and A. Praktiknjo, "How to model European electricity load profiles using artificial neural networks," *Applied Energy*, vol. 277, Nov. 2020, doi: 10.1016/j.apenergy.2020.115564.
- [38] J. Son, J. Cha, H. Kim, and Y. M. Wi, "Day-Ahead Short-Term Load Forecasting for Holidays Based on Modification of Similar Days' Load Profiles," *IEEE Access*, vol. 10, pp. 17864–17880, 2022, doi: 10.1109/ACCESS.2022.3150344.
- [39] C. Dominguez, K. Orehounig, and J. Carmeliet, "Estimating hourly lighting load profiles of rural households in East Africa applying a data-driven characterization of occupant behavior and lighting devices ownership," *Development Engineering*, vol. 6, Jan. 2021, doi: 10.1016/j.deveng.2021.100073.
- [40] B. Gao, X. Liu, and Z. Zhu, "A bottom-up model for household load profile based on the consumption behavior of residents," *Energies (Basel)*, vol. 11, no. 8, 2018, doi: 10.3390/en11082112.
- [41] A. Marszal-Pomianowska, P. Heiselberg, and O. Kalyanova Larsen, "Household electricity demand profiles - A high-resolution load model to facilitate modelling of energy flexible buildings," *Energy*, vol. 103, pp. 487–501, May 2016, doi: 10.1016/j.energy.2016.02.159.

- [42] Z. Wang and T. Hong, "Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN)," *Energy and Buildings*, vol. 224, Oct. 2020, doi: 10.1016/j.enbuild.2020.110299.
- [43] J. Z. Kolter and M. J. Johnson, "REDD: A Public Data Set for Energy Disaggregation Research." [Online]. Available: <http://www.enmetric.com>
- [44] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study," *Scientific Data*, vol. 4, Jan. 2017, doi: 10.1038/sdata.2016.122.
- [45] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study," *Scientific Data*, vol. 4, Jan. 2017, doi: 10.1038/sdata.2016.122.
- [46] M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, "A Topic-Based Hidden Markov Model for Real-Time Spam Tweets Filtering," in *Procedia Computer Science*, 2017, vol. 112, pp. 833–843. doi: 10.1016/j.procs.2017.08.075.
- [47] G. Y. Lin, F. Pan, Y. Y. Yang, L. Yang, G. He, and S. Fan, "The Pattern Recognition of Residential Power Consumption Based on HMM," in *International Conference on Innovative Smart Grid Technologies, ISGT Asia 2018*, Sep. 2018, pp. 413–418. doi: 10.1109/ISGT-Asia.2018.8467905.
- [48] A. Munther, R. R. Othman, A. S. Alsaadi, and M. Anbar, "A performance study of hidden markov model and random forest in internet traffic classification," in *Lecture Notes in Electrical Engineering*, 2016, vol. 376, pp. 319–329. doi: 10.1007/978-981-10-0557-2\_32.
- [49] N. D. Hoang and D. Tien Bui, "Slope Stability Evaluation Using Radial Basis Function Neural Network, Least Squares Support Vector Machines, and Extreme Learning Machine," in *Handbook of Neural Computation*, 2017. doi: 10.1016/B978-0-12-811318-9.00018-1.
- [50] A. Arif, T. A. Alghamdi, Z. A. Khan, and N. Javaid, "Towards Efficient Energy Utilization Using Big Data Analytics in Smart Cities for Electricity Theft Detection," *Big Data Research*, vol. 27, Feb. 2022, doi: 10.1016/j.bdr.2021.100285.