REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

PREDICTING AND ANALYSIS ELECTRICAL ENERGY CONSUMPTION BY USING DATA MINING ALGORITHMS

Mohamed Abdulhamed Mohamed ALMUSLEHI

Master's Thesis

Supervisor

Asst. Prof. Dr. Sefer KURNAZ

Istanbul, 2023

PREDICTING AND ANALYSIS ELECTRICAL ENERGY CONSUMPTION BY USING DATA MINING ALGORITHMS

Mohamed Abdulhamed Mohamed ALMUSLEHI

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled AN PREDICTING AND ANALYSIS ELECTRICAL ENERGY CONSUMPTION BY USING DATA MINING ALGORITHMS prepared by MOHAMED ABDELHAMID MOHAMED ALMUSLEHI and submitted on 08/08/2023 has been **accepted unanimously** for the degree Master of Science in Electrical and Computer Engineering.

____

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Thesis Defense Jury Members:

Asst. Prof. Dr. Sefer KURNAZ	Department of computer Engineering, Altınbas University	__________________
Asst. Prof. Dr. Oguz KARAN	Department of Software Engineering, Altınbas University	__________________
Asst. Prof. Dr. Zeynep ALTAN	Department of Software Engineering, Beykent University	__________________

I hereby declare that this thesis meets all format and submission requirements of a Master's Thesis.

Submission date of the thesis to the Institute of Graduate Studies: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Mohamed Abdulhamed Mohamed ALMUSLEHI

Signature

DEDICATION

I devote and pledge this research work to my supervisor who is salient for guiding me through whole research work as well as my family for always assisting me in my hard time.

PREFACE

First and foremost, I would like to thank my supervisor Asst. Prof. Dr. Sefer KURNAZ for guiding and helping me along the way in writing this dissertation. Discussing my progress, problems, and ideas with my supervisor Asst. Prof. Dr. Sefer KURNAZ a couple of times every week helped me tremendously in understanding the logic behind the research. It made me better realize the technical need for this research work.

ABSTRACT

PREDICTING AND ANALYSIS ELECTRICAL ENERGY CONSUMPTION BY USING DATA MINING ALGORITHMS

ALMUSLEHI, Mohamed Abdulhamed Mohamed
M.Sc., Electrical and Computer Engineering, Altınbaş University,
Supervisor: Asst. Prof. Dr. Sefer KURNAZ
Date: 08/2023
Pages: 59

This study accumulates and presents pertinent data in a variety of formats for the purpose of classifying power consumption by activity. It has been fascinating to determine how to graphically depict each stage and classify electrical data in a way that satisfies all requirements. This facilitates the detection of issues and anomalies and simplifies the process of comparing electricity consumption. In the coming decades, electricity will continue to gain prominence as a primary energy source. Smart circuits and smart meters offer numerous benefits to both the utility and the customer. This study combined classification (specifically five algorithms) and clustering theory, with energy consumption per hour (%) functioning as the common framework, in order to classify electricity use based on the similarities of electrical load profiles. After classifying everyone, we will be able to offer each subset advice on how to save money and energy. Consequently, individuals will be more aware of their electricity consumption and motivated to take steps to reduce it. A post-clustering and classification study that uses Weka for analysis and result generation employs an iterative technique based on computational classification calculation to identify anomalies and reallocate them to more acceptable classes. When compared, Decision Tree, Support Vector Machine, Naive Bayes, Random Forest, and Hybrid are the five classification techniques that yield identical results. Categorizing power consumption permits a deeper understanding of the relationship between human behavior and electricity consumption. It improves the quality of the energy conservation consulting service and the customer experience by providing timely, relevant advice based on the unique characteristics of each individual consumer.

Keywords: Prediction, Electricity Consumption, Smart Meter, Classification, Clustering, Intelligent System, Data Mining, Electric Usage.

TABLE OF CONTENTS

Pages

ABSTRACT .... **vii**

LIST OF TABLES .... **xii**

LIST OF FIGURES .... **xiii**

ABBREVIATIONS .... **xvi**

1. INTRODUCTION 1

1.1 ENERGY AND SUSTAINABLE DEVELOPMENT 1

1.2 ENERGY SUPPLY & DEMAND 4

1.3 ENERGY POLICIES AND ENERGY EFFICIENCY IN EUROPE 5

1.3.1 European Energy Strategy 6

1.3.2 Energy Efficiency 6

1.3.3 Active Consumer 7

2. LITERATURE REVIEW **9**

2.1 SMART GRIDS 10

2.1.1 Definition 10

2.1.2 Smart Grid Characteristics, Functions, Benefits 10

2.1.3 Smart Grid Conceptual Model 12

2.2 SMART METERS & ADVANED METERING INFRASTRUCTURE 16

2.2.1 Advanced Metering Infrastructure 17

2.2.2 Smart Meters 22

3. METHODOLOGY **25**

3.1 OBJECTIVE OF CLEANING DATA 25

3.2 ELECRICITY CONSUMOTION DATA 26

3.3 DATA EXPLORATION AND VISULIZATION 27

3.4 OBJECTIVE OF PREDICTION 28

3.5 ENERGY CONSUMPTION VISUALIZATIONS 28

3.6 CLASSIFICSTION OBJECTIVE 29

3.7 CLUSTERING OVERVIEW 29

3.7.1 Cluster Analysis Theory 29

3.7.2 Clustering and Classification 30

3.7.3 Times Series Clustering and Classification 31

3.7.4 Distances and Linkages 32

3.8 CLASSIFICATION ALGORITHM 33

3.8.1 Support Vector Machines 33

3.8.2 Random Forest 34

3.8.3 Decision Tree 34

3.8.4 Naive Bayes 35

3.8.5 Hybrid Model 36

4. RESULTS 37

4.1 FALSE POSITIVE AND FALSE NEGATIVE RATE 37

4.2 TRUE POSITIVE AND TURE NEGATIVE RATE 38

4.3 CLASSIFICATION ON STATISTICAL DATASET 38

4.3.1 Decision Tree Prediction 39

4.3.2 Random Forest Prediction 39

4.3.3 Support Vector Machine Prediction 40

4.3.4 Naive Bayes Prediction 40

4.3.5 Hybrid Model Prediction 41

4.3.6 Prediction Result 42

5. CONCLUSION 43

5.1 FUTURE WORK 44

REFERENCES 46

LIST OF TABLES

Pages

Table 3.1: Basic Characteristics to Differentiate Clustering and Classification.................. 31

Table 3.2: Summary of the Clustering and Classification Techniques and Papers, Adapted From [35].. 31

LIST OF FIGURES

Pages

Figure 1.1: Energy and Sustainable Development Linkages[5]. 1

Figure 1.2: Correlation GDP vs Energy Consumption. Energy Consumption Per Capita in thousands of Tons of Equivalent Oil (toe) and Gross Domestic Product Per Capita in million $[5]. 2

Figure 1.3: Energy Use Per Unit of GDP Trend. [7]. 3

Figure 1.4: Energy Use Per Country in kWh vs Human Development Index[7]. 3

Figure 1.5: Forecast of Energy Demand by Non-OECD Countries[11]. 4

Figure 1.6: Forecast of Energy Demand on OECD Countries by Type of Fuel [11]. 5

Figure 1.7: Comparison Energy Consumption by Sector Shares and Total from 1973 (Left) to 2011 (right)[11]. 5

Figure 1.8: Centralized and Decentralized Conceptual Models are Contrasted [17]. 8

Figure 2.1: Electricity Consumption Expected % Increase by Regions. [19]. 9

Figure 2.2: Smarter Electricity Systems, from Past to Future [20]. 10

Figure 2.3: Conceptual Illustration with all Domains Participating in a Smart Grid[21]. 12

Figure 2.4: Deep Look at the Bulk Generation Domain[23]. 13

Figure 2.5: Deep Look at the Customer Domain [21]. 14

Figure 2.6: Deep Lok at the Market Domain. 15

Figure 2.7: Deep Look at the Servicer Provider Domain. [21]. 16

Figure 2.8: Advanced Metering Infrastructure, from Customer's Home to Stakeholder's Services [21]. 18

Figure 2.9: Monitoring Energy Consumption in Different Supports[23]. 18

Figure 2.10: Conceptual Model Diagram of a Home Area Network (HAN) and Features [25]. 19

Figure 2.11: European Smart Meter with Open Smart Grid Protocol[25]. 19

Figure 2.12: Wide Area Network (WAN) Model, Communicating and Aggregating Various Smart Meter Measures from the Home and Local Area Networks [25]. 20

Figure 2.13: Meter Data Management System Inputs/Outputs and Information Systems Served[27].. 21

Figure 2.14: Time of Use Tariffs Graphs, Electricity is More Expensive in the Afternoon where the Peak Demand [23]. 22

Figure 2.15: Scheme of Net Metering, Two Way Smart Meter Importing and Exporting Electricity with the Grid [29]. 23

Figure 3.1: States of the Data When Performing a Data Analysis [25].............................. 26

Figure 3.2: Diagram to Explain How the Data Used for the Analysis is Obtained [27]. 27

Figure 3.3: Steps to Perform Any Cluster Analysis 30

Figure 3.4: Illustration of the Euclidean and Manhattan Distances [45]............................. 30

Figure 3.5: Example a Two-Class Linear SVM Classifier [24]. 33

Figure 3.6: Phases of Ensemble Learning Approaches to Solve Classification Problems [31] …..…………………………………………………………………………………..…34

Figure 3.7: Decision Tree Algorithm Structure [34]. 35

Figure 3.8: Naive Bayes Algorithm on Left with Respect to Support Vector Machine on Right side for Classification Structure [34]. 35

Figure 3.9: Hybrid Model Algorithm Present as Combination of Random Forest and Decision Tree for Classification Structure [4]. 36

Figure 4.1: The Overall Classification Results Based on Approaches Used on Statistical Dataset. 39

Figure 4.2: The Decision Tree Classification Results Based on Approaches Used. 39

Figure 4.3: The Random Forest Classification Results Based on Approaches Used. 40

Figure 4.4: The Support Vector Machine Classification Results Based on Approaches Used ………………………………………………………………………………………..40

Figure 4.5: The Naive Bayes Classification Results Based on Approaches Used. 41

Figure 4.6: The Hybrid Model Classification Results Based on Approaches Used............ 41

Figure 4.7: The Prediction of Result for Classifying the Consumption of Electricity in Terms of Mean Square Error. 42

ABBREVIATIONS

AMI	:	Advanced Metering Infrastructure
IEA	:	International Energy Agency's
HDI	:	Human Development Index
CHP	:	Combined Heat and Power
DLI	:	Dynamic Line Rating
AMP	:	Advanced Malware Protection
NGIPS	:	Next-Generation Intrusion Prevention System
ML	:	Machine Learning

1. INTRODUCTION

1.1 ENERGY AND SUSTAINABLE DEVELOPMENT

The expanding significance of energy means that its importance to the advancement of human civilization will increase. Adjusting to altering customer requirements and regulatory constraints presents significant challenges to the contemporary energy sector. Recently, the policy community has paid considerable attention to environmental issues [1]. Three key international environmental conferences, Stockholm 1972, Rio 1992, and Johannesburg 2002, offer proof[2]. From these discussions arises sustainable development, an all-encompassing strategy that simultaneously addresses economic, social, and environmental challenges [3]. Figure 1.1 illustrates the importance of energy to all sustainable development concepts [4-5].

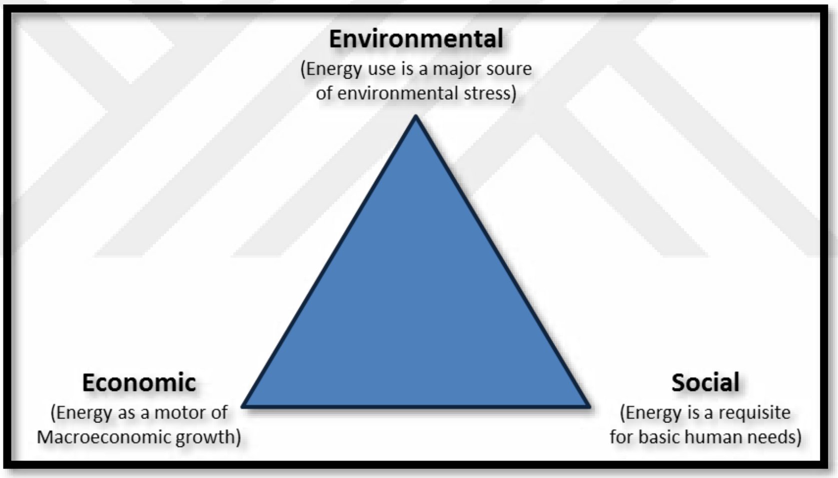


Figure 1.1: Energy and Sustainable Development Linkages [5].

Historically, conventional energy sources, primarily fossil fuels such as oil, coal, and natural gas, have provided abundant and affordable energy through their extraction, refining, and consumption. Nonetheless, pervasive, regional, and local exploitation of finite, nonrenewable resources has resulted in extensive environmental damage. This causes climate change, acid deposition, and urban pollution [6].

Solar (photovoltaic and thermal), wind (wind and biomass), and hydropower (hydroelectric) are examples of renewable energy sources. This energy's production, conversion, and use have negligible environmental impacts, reducing the rate of global warming.

Figure 1.2 depicts the relationship between energy and GDP (a macroeconomic metric).

Figure 1.2 depicts the relationship between energy and GDP (a macroeconomic metric). Industrial operations that convert basic materials into finished products or services frequently rely on the use of energy [7].

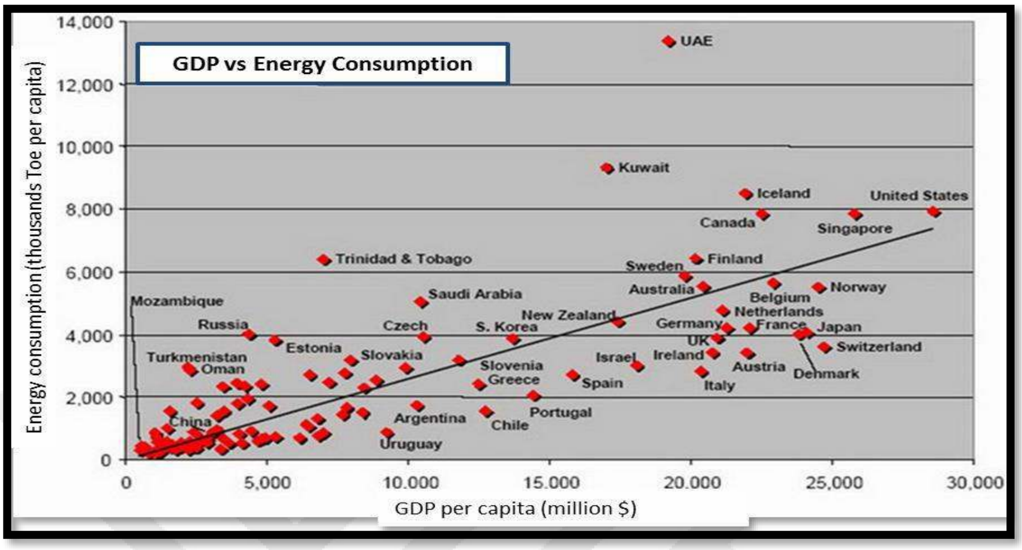


Figure 1.2: Correlation GDP vs Energy consumption. Energy Consumption Per Capita in Thousands of Tons of Equivalent Oil (Toe) and Gross Domestic Product Per Capita in Million $ [5].

The preceding link may appear obvious, but the following two pieces of evidence are even more compelling: Both industrialized and emerging nations are decreasing their energy intensity over time, as shown in Figure 1.3. Energy intensity is calculated by dividing total energy consumption by GDP.

Energy efficiency is the polar opposite of energy intensity for a given economy. Among the top 40 economies, there are two distinct categories: developed economies, with a high GDP per capita but low efficiency, and developing economies, with a low GDP per capita and high inefficiency. [7].

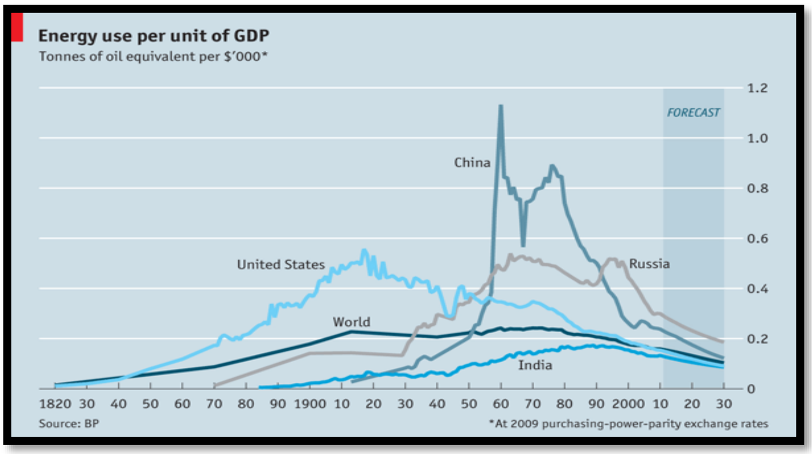


Figure 1.3: Energy Use Per Unit of GDP Trend. [7].

Reducing energy consumption while maintaining or increasing economic output (also known as "decoupling" energy consumption from economic growth [5]) is a recently popular concept. The Energy Transition in Germany could stand to be more energy efficient.

The more intricate energy link addresses the social aspect of sustainable development by providing people with their fundamental needs. All of our fundamental requirements for sustenance, protection, comfort, health, education, and employment require energy. The Human Development Index (HDI) of a country is determined using variables such as per capita income, life expectancy, and educational attainment. It was created so that countries' human development progress could be compared and contrasted. In its Human Development Reports, the UNDP publishes each year the data used to calculate the HDI and equivalent indices.

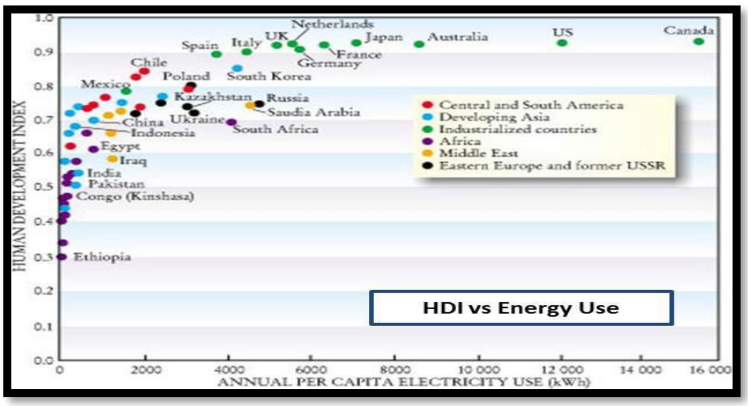


Figure 1.4: Energy use Per Country in kWh vs Human Development Index [7].

Numerous European studies and publications, including the 2014 Energy Supply and demand study by the European Commission, have demonstrated the strong correlation between socioeconomic indicators and energy consumption. Comparing household energy consumption, developing nations consume substantially less energy than developed nations [8].

Energy poverty (defined as a lack of access to electricity or cooking facilities) is largely caused by inadequate access to energy services. According to the International Energy Agency, only about 18% of individuals have access to electricity (IEA: Energy Poverty, 2015). However, energy poverty is not limited to developing nations; as a consequence of the global economic crisis and rising energy costs, a number of middle- and upper-class families in developed countries have recently found themselves in a similar situation [9].

1.2 ENERGY SUPPLY & DEMAND

Energy consumption is projected to increase over the next two decades, possibly tripling by 2035. The majority of growth will originate from developing and non-OECD countries, particularly China and India. Figure 1.5 illustrates the International Energy Agency's (IEA) forecast for 2035 [10], based on information from the IEA's 2012 Global Energy Outlook report.

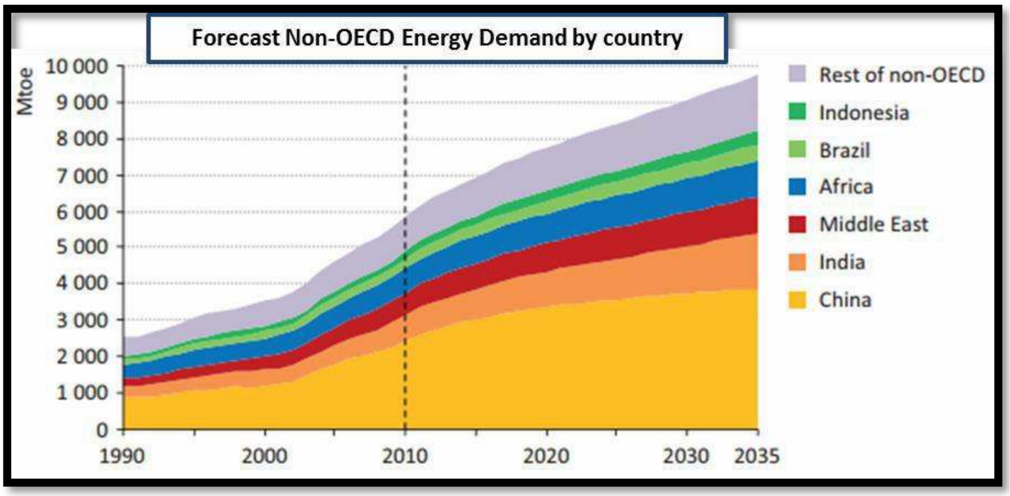


Figure 1.5: Forecast of Energy Demand by non-OECD Countries [11].

Energy strategies frequently seek to increase either energy efficiency or the use of renewable energy sources. By reducing carbon dioxide emissions, improvements in energy efficiency and the use of renewable energy sources will contribute to the decarbonization of the economy. Figure 1.6 depicts the timeline for the complete substitution of fossil fuels with renewable energy sources. We will have to rely on tried-and-true primary energy sources to meet the rising demand [11].

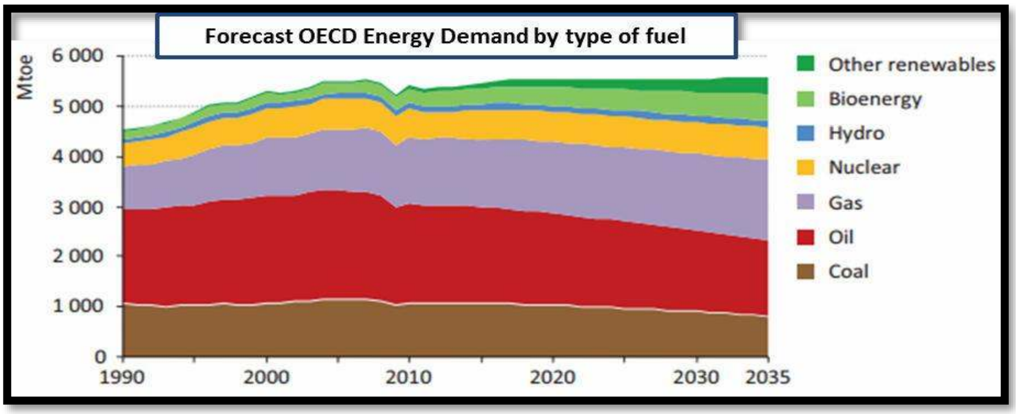


Figure 1.6: Forecast of Energy Demand on OECD Countries by Type of Fuel [11].

Figure 1.7 depicts the sectoral energy consumption patterns. Compared to 1973, when the proportions were relatively stable, the energy consumption in 2011 was nearly twice as high. If we have a greater understanding of where and how this energy is used, we can create more effective laws for each industry [11].

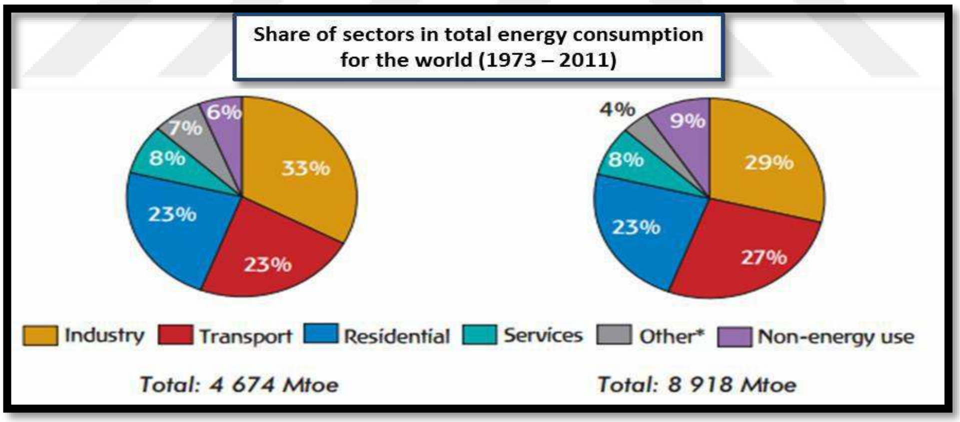


Figure 1.7: Comparison Energy Consumption By Sector Shares and Total From 1973 (Left) to 2011 (Right) [11].

1.3 ENERGY POLICIES AND ENERGY EFFICIENCY IN EUROPE

In an effort to make energy more reliable, accessible, and environmentally friendly, the European Commission has designated the formation of an Energy Union as one of the top ten objectives for the European Union.

1.3.1 European Energy Strategy

The European Union is the largest importer of energy in the world, importing 90% of its petroleum oil and 66% of its natural gas (European Commission: Imports and Secure Supply, 2015). This equates to annual expenditures of over 400 billion euros for these nations. Energy Union, a 2015 report published by the European Commission, found that 12 of its member states were "energy-isolated islands" with only 10% connection, and that 6 of those were dependent on a single supplier, leaving their consumers more vulnerable to price fluctuations and instability [12]. The rising cost of conducting business in Europe is illustrated by the possibility that gasoline prices will treble and that energy prices will be 30 percent higher than in the United States. Standardizing market regulations and liberalizing markets are potential solutions. Once integration is complete, it is anticipated that annual savings of up to 40 billion euros will be realized as a consequence of reduced reliance on foreign suppliers. Due to the fact that more than 10% of Europeans are unable to pay their energy expenses because of the economic crisis, it is impossible to ensure that all EU citizens have access to electricity [13].

1.3.2 Energy Efficiency

According to The Economist (2015), the best strategy is to optimize energy efficiency, also known as the "invisible fuel." The energy industry has now acknowledged the significance of energy efficiency. Financially and environmentally, the entire energy supply chain must be optimized, particularly for countries like Europe that rely so heavily on imported energy.
The regulations and mandates of the Energy Efficiency Directive are geared toward this objective. The European Union intends to reduce its energy consumption by 20 percent by the year 2020, aiming to use no more than 1,474 terajoules of fundamental energy and 1,078 terajoules of final energy.

Energy efficiency in the energy supply focuses primarily on the implementation of district heating and cooling systems, the use of renewable energy sources, and the development of highly efficient cogeneration facilities (that generate both electricity and heat). Grid losses (accounting for nearly 30%) resulting from energy generation, transit, and distribution, as well as the energy savings attained as a result of suppliers' obligations and the use of white certificates, are the ultimate drivers of smart grid implementation (Suppliers Obligations & White Certificates, 2015).

The focus of this article is Chapter II of the Energy Efficiency Directive. By 2020, 72% of all electricity will be measured by smart meters, allowing access to both historical and real-time usage and invoicing information. Customers will then have the knowledge necessary to make informed decisions regarding their energy consumption [15].

In addition to the Energy Efficiency Directive, three additional directives address the thermal and electrical demands of structures, which consume more than 40 percent of all energy used in Europe (European Commission: Energy Efficiency, 2015):

1.3.3 Active Consumer

There appears to have been a paradigm shift in the energy industry, as new regulations are being implemented to safeguard consumers' rights and ability to receive adequate care. The next two items on the list are crucial to the mission at hand, which is to monitor and control energy consumption:

A consumer has the legal right to receive accurate information regarding their monthly gas and electricity expenses. It is necessary to educate consumers on energy consumption and conservation methods. The potential of renewable energy as a dependable source of energy for transportation and industry.

Policymakers are also pursuing the modernization of new energy infrastructure through the incorporation of innovation and cutting-edge technologies.

As shown in Figure 1.8, the paradigm for energy production is swiftly shifting from centralized to decentralized production.

Due to factors such as distance from industrial locations, a one-way network, and stringent energy consumption [16], customers in this scenario act primarily as bystanders. Few extremely large power facilities are used for central electricity generation.

As opposed to merely providing a service, the new decentralized manufacturing model includes enhancements that benefit the entire process. A two-way network makes it possible to store energy for use when and where it is required, reduce consumption peaks through the use of information and communication technology and big data, and engage the customer more directly. Using advanced technologies such as fuel cells [13] or other energy sources (such as photovoltaics [PV], wind, etc.), small and medium-sized power facilities are able to relocate production closer to the point of consumption.

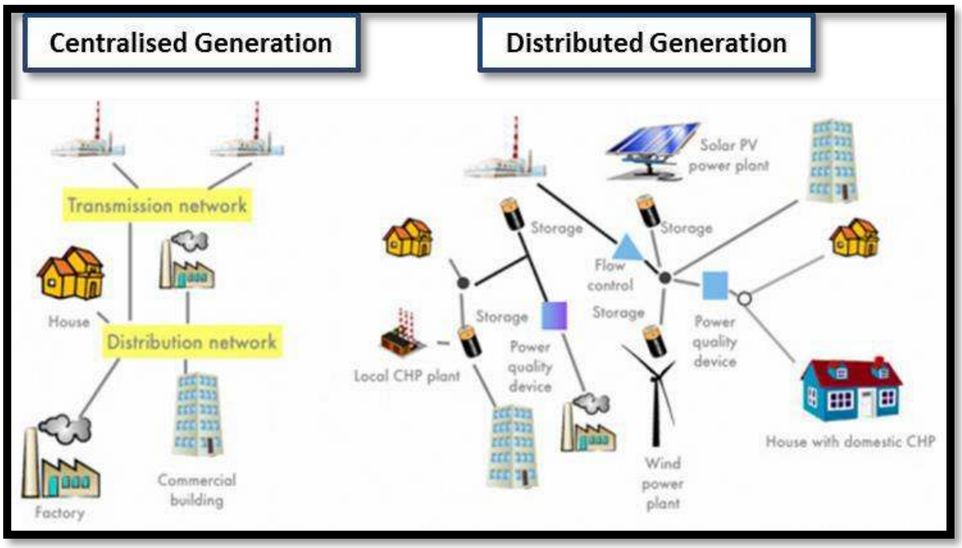


Figure 1.8: Centralized and Decentralized Conceptual Models are Contrasted [17].

Numerous studies, such as [18], predict that almost all electronics, automobiles, and other equipment will be electrically powered in the near future. Therefore, smart grids are essential for the modernization of the electrical infrastructure in order to meet expanding demand.

2. LITERATURE REVIEW

Electricity consumption is expected to increase quicker than any other energy source by 2050 [19], constituting a significant portion of the market. As depicted in Figure 2.1, this expansion will be led by developing and emerging economies, with some established nations also achieving modest growth.

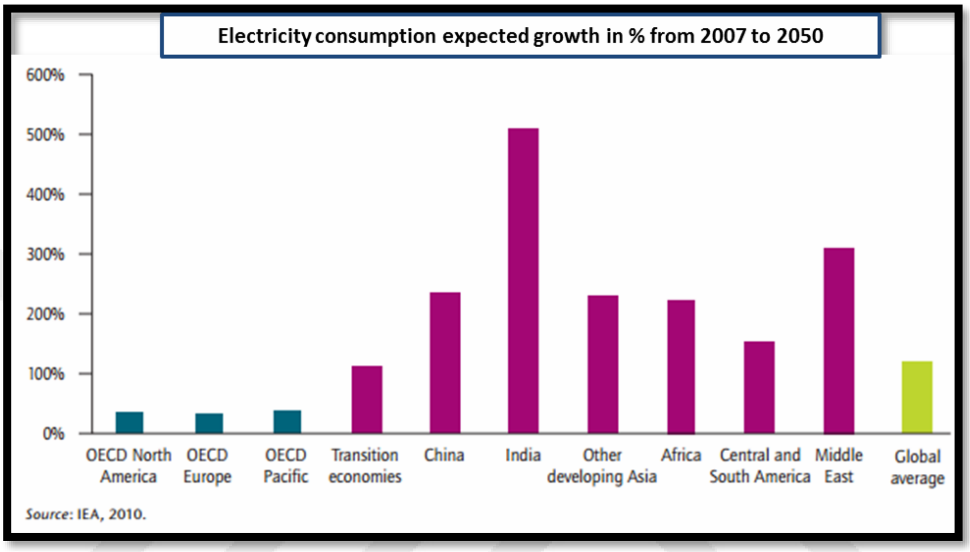


Figure 2.1: Electricity Consumption Expected % Increase by Regions. [19].

Many individuals believe heat pumps are to blame for the significant increase in energy consumption by residential heating and conditioning systems. Large-scale turbines that can cool or heat an entire neighborhood.

The pervasive adoption of plug-in hybrid and all-electric vehicles has contributed significantly to the electrification of transportation and mobility.

CHP (combined heat and power), biomass, solar, wind, and hydropower are all examples of variable electricity generators.

This transition will mitigate the effects of climate change while allowing energy sources that emit fewer greenhouse gases to meet the same or greater demand. However, economic and social development cannot continue until the electrical system is updated to accommodate these innovative features.

2.1 SMART GRIDS

2.1.1 Definition

Despite its extensive application, the term "smart grid" remains undefined. Consider it a modification and upgrade of the existing electrical system to satisfy present and future needs.

The objective of "smart grids" is to improve the power grid in terms of efficiency, reliability, and safety while reducing costs and environmental impact using cutting-edge technologies such as distributed electricity producers, electricity storage, and electric vehicles, among others. Providing power generators and consumers with the means to encourage flexible output and proactive participation. Figure 2.2 illustrates that the effectiveness of the new electric grid will depend significantly on its customers.

Clearly distinguishable from the smart grid's centralized control centers that supervise electricity transmission and distribution are power plants and consumers [19].

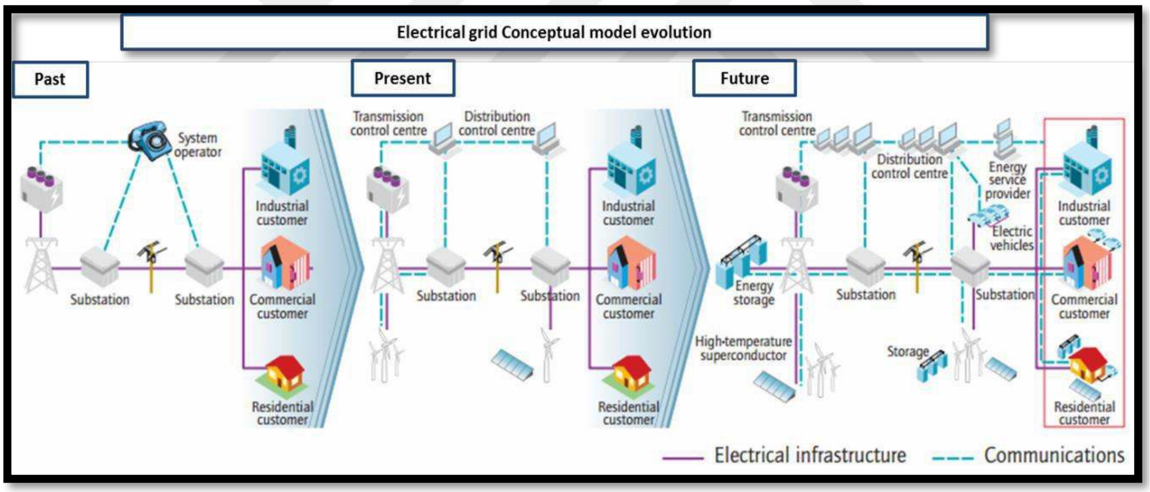


Figure 2.2: Smarter Electricity Systems, From Past to Future [20].

2.1.2 Smart Grid Characteristics, Functions, Benefits

a. Reliability, quality, and security

Today's grid is equipped with cutting-edge technologies for operating transmission systems, real-time monitoring of grid health, and self-assessments to locate, evaluate, and rectify faulty grid components.

By autonomously detecting grid interruptions and responding to unanticipated failures or attacks, the system enables the remainder of the affected zone to resume normal operations by isolating and preventing problem areas. Self-healing actions taken by the grid in response to fewer disruptions will enhance service quality and permit more effective management of the delivery infrastructure.

b. Ability to accommodate all types of generation and storage options:

Smart grids facilitate the coordination of dispersed power plants of differing technological sophistication and size. The new infrastructure permits the connection of any "plug and play" generating facility, such as rooftop solar photovoltaic arrays, combined heat and power plants, and medium-scale wind farms, to the grid.

The addition of energy storage (batteries, hydrogen, compressed air, pumped hydro, etc.) to the grid enhances its performance and reduces its operating expenses. To accommodate the growing number of electric vehicles, the grid could serve as a storage system.

c. Consumer engagement:

Since consumers will be able to track their expenditures and consumption in real time, they will be in a strong position to influence future electric grid developments. In response to new power pricing models, such as time-of-use tariffs, which increase the price of energy during peak hours, consumers can assist the environment and save money on electricity by changing their behavior.

In the future, the grid will be able to limit client consumption autonomously when electricity is expensive or scarce. Until then, turning off appliances may incur fees from utility providers.

d. Optimize assets utilization and operating efficiency.

Distributors can pinpoint the source of power outages and allocate resources accordingly, saving money on preventative maintenance and emergency restorations, with increased network visibility.

"Smart grids" maximize the effectiveness of their power distribution networks by integrating cutting-edge technological advancements. The objective of the control devices is the efficient operation of the system; therefore, they will make adjustments to the operation to achieve this.

e. Enable new products, services, and markets.

Markets can influence energy, capacity, location, timing, and electrical quality, making them an integral component of grid management. As a result of deregulation's open-access (liberalization) market, consumers will have greater access to a variety of services.

2.1.3 Smart Grid Conceptual Model

The National Institute of Standards and Technology devised a conceptual model of a "smart grid" [22] to help readers comprehend what a "smart grid" is and how it may be subdivided into multiple sets. Figure 2.3 illustrates seven distinct categories, including "bulk production," "transmission," "distribution," "customer," "markets," "operators," and "service providers".

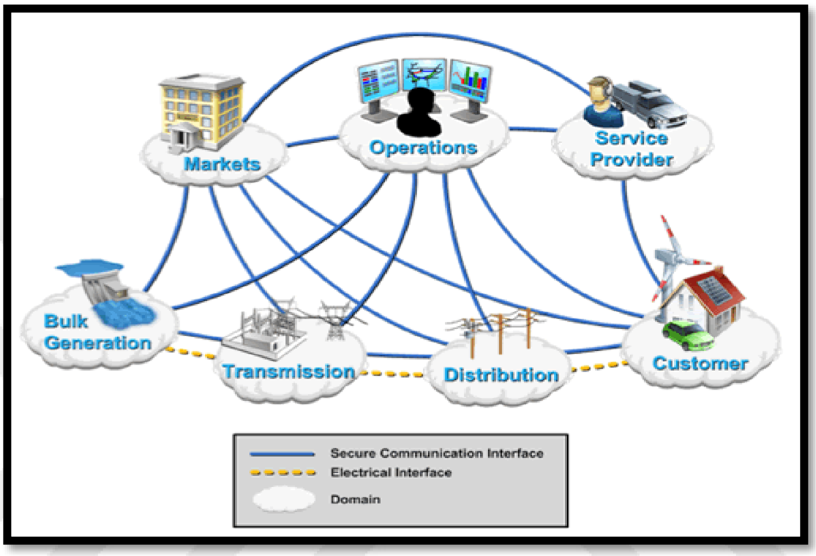


Figure 2.3: Conceptual Illustration with All Domains Participating in A Smart Grid [21].

a. Bulk generation:

As depicted in Figure 2.4, a diverse array of energy sources and technological scales can be utilized to power large-scale, or "bulk," electrical infrastructure.

Coal, natural gas, and nuclear fusion power facilities are all conventional, nonrenewable, and unsustainable energy sources. Hydroelectric, biomass, geothermal, and pump storage are examples of dependable and durable energy sources. Solar and wind energy are examples of adaptable and renewable energy sources.

The bulk generating domain can begin exchanging data with the operations, markets, and transmission domains once power has been generated and transmitted. The ICT infrastructure enables the grid to autonomously reroute electricity flow when specific generators fail. Utilizing energy storage devices, starting and stopping power facilities, and distributing or storing energy are all viable options.

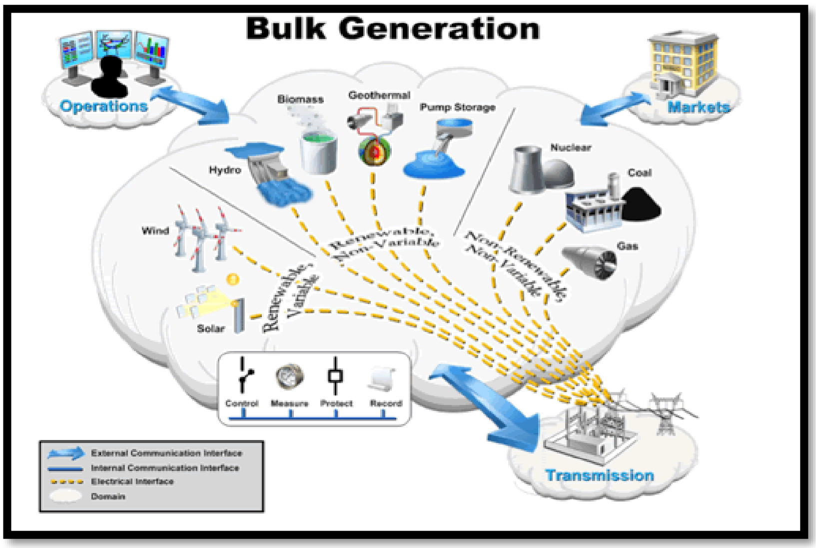


Figure 2.4: Deep look at the Bulk Generation Domain [23].

b. Transmission:

The network of substations that transports electricity from its source to its destination is known as transmission. It is wired to the regions of generation and distribution and communicates with the areas of operations and markets. To maintain grid reliability, transmission network operators must establish a balance between supply and demand. These sources, which may be distributed, may include storage and peaking generating units, which enable more accurate energy control and distribution.

a. Furthermore, required are physical technologies, such as those utilized to improve the transmission system.
b. Using sensors to monitor a network segment's carrying capacity, the Dynamic Line Rating (DLR) maximizes transmission assets while introducing an overflow risk.
c. Using flexible AC transmission systems increases the power transmission capacity and enhances transmission network control.
d. With high voltage DC, system losses can be decreased by linking offshore wind or solar farms to load centers.
e. By decreasing transmission losses, high-temperature superconductors improve transmission efficiency.

c. Distribution:

The Transmission domain communicates electronically with the Consumer domain, whereas the Distribution domain communicates with the Markets and Operations domains. Distribution grids have changed dramatically over time, transitioning from hierarchical, unidirectional interfaces with limited communication to bidirectional grids with numerous monitoring and control devices that can disperse distributed generation and storage, regulate demand response and the load, and improve reliability.

Distribution grid management enhances the intelligence of the grid at distribution and substations by automating distribution procedures and utilizing real-time data for problem detection and rapid restoration.

d. Customer:

To be fully functional, the future smart grid will require the participation of all of its consumers. Due to the grid, everyone who wishes to contribute to the writing of this essay has access to the necessary materials. display Figure 2.5 to illustrate the relationship between the Distribution and Consumer domains, as well as the Operations, Markets, and Service Providers domains.

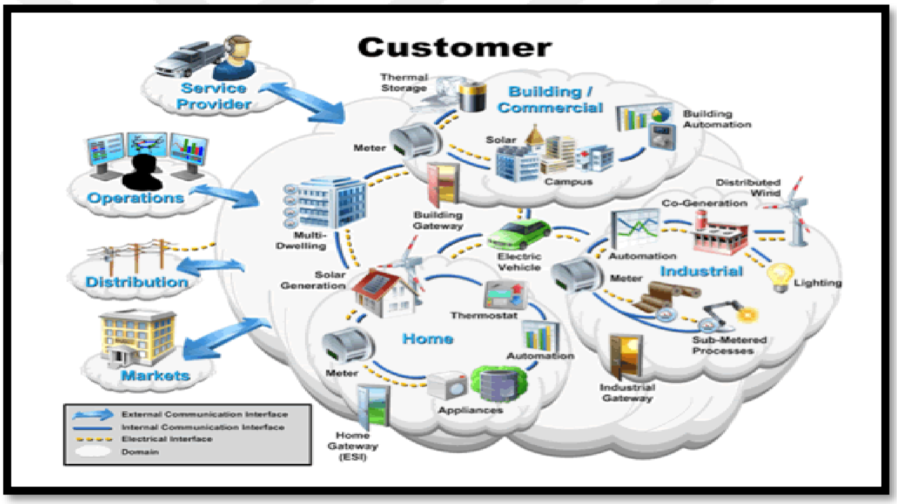


Figure 2.5: Deep Look at the Customer Domain [21].

Customers can manage their energy consumption and production due to the bidirectional flow of information and commands in a smart grid. Despite the vast differences in size, customers in the industrial, commercial/building, and residential sectors all have access to the same state-of-the-art metering system, vehicle charging infrastructure, and customer-side devices.

Advanced metering infrastructure aims to enable utilities and consumers to communicate in real time regarding energy price and consumption data (such as time of day and total amount of electricity used). AMI meter data management system incorporates interaction with BMS, in-home consumption displays, monitoring and controlling distributed generation, and remote load management.

It is anticipated that the EV charging infrastructure, which consists of chargers, batteries, and inverters, will offer capacity reserve and peak load reduction capabilities. By providing a place to store excess energy during periods of low demand, intelligent charging and discharging can have a positive impact on electricity costs and expenses. requires bidirectional communication between the AMI and the customer's infrastructure.

Smart appliances, interior displays, energy management systems, and dashboards are customer-side solutions that can be used to monitor and control electrical use in the home, workplace, and factory. Increase energy efficiency and decrease peak demand with this technology, consumer-managed demand response, and even automated and remotely controlled appliances that react to price changes.

The markets are responsible for ensuring pricing and supply and demand balance in the system. As indicated in Figure 2.6, this system matching relies extensively on interactions with the generating, transmission, distribution, and customer domains.

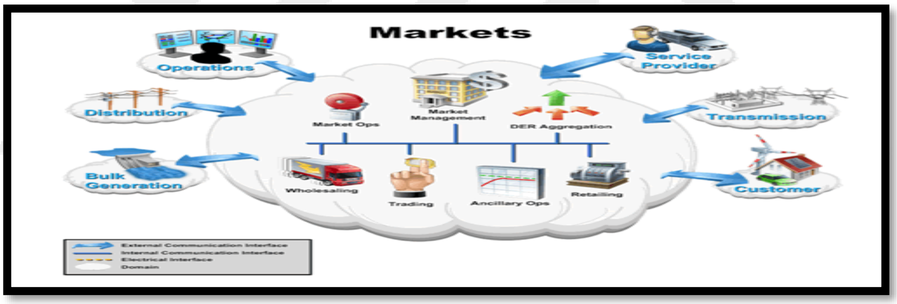


Figure 2.6: Deep Look at the Market Domain [21].

Operators conduct constant management tasks, such as network operation, monitoring, control, fault management, operation feedback analysis, operational statistics and reporting, real-time network computation, and dispatcher training.

They are responsible for ensuring the efficient operation of the electrical system. Among the responsibilities of the regulated utility are maintenance and construction, meter reading, and security management.

Service providers are businesses that provide services to utilities and electrical clients as well as operational support to power system manufacturers, distributors, and end users. The Service Provider and the Markets, Operations, and Customers sectors engage in an open interaction.

The interaction between the market and consumer domains will permit the formation of unique services that may be utilized to satisfy the market's demands and desires (see Figure 2.7). The demand for new services may be filled by electric service providers or third parties.

Alongside tried-and-true solutions such as billing and client account management, new services such as home energy management or generation can be provided. Software as a service (SaaS) will have a significant impact on this industry by providing users with detailed and user-friendly data on their electricity bills and usage.

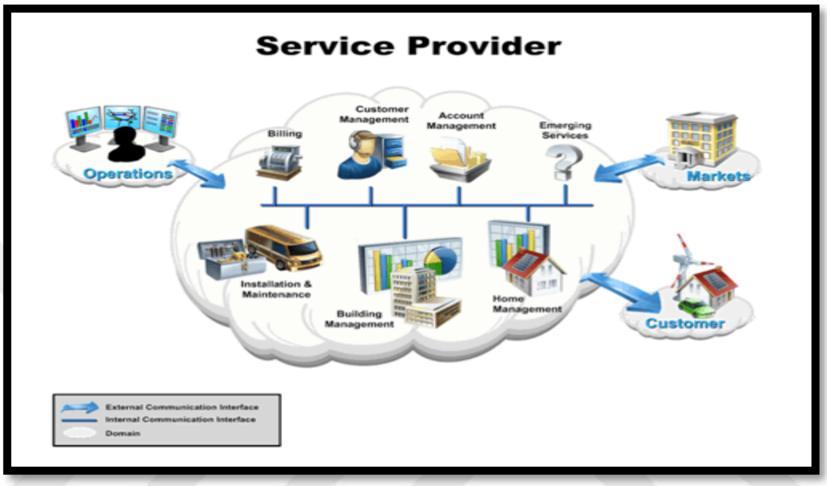


Figure 2.7: Deep Look at the Servicer Provider Domain. [21].

2.2 SMART METERS & ADVANED METERING INFRASTRUCTURE

Smart meters and the Advanced Metering Infrastructure (AMI) are given serious consideration because they supply the consumption data used in this inquiry. Even though this infrastructure is still under construction, a number of governing bodies and regulators have set their eyes on implementing smart meters. Within the context of the Energy Policy 2020 program, the European Union acts as a model.

Due to stringent government regulation, the electrical markets of other nations were reluctant to absorb new technology and adjust to fluctuating market conditions. Private smart meters have been widely implemented due to the long-term financial benefits of energy management, particularly in commercial, industrial, and office buildings. Unfortunately, in the domestic sector, the energy savings do not cover the cost of private submetering equipment. So, many are awaiting the "official" adoption of smart meters before beginning to restrict their energy consumption.

2.2.1 Advanced Metering Infrastructure

What is AMI?

By merging numerous technologies into a single framework, AMI enables utilities to communicate with their customers in both directions. These technologies make possible measurements, data collection, and real-time transmission to grid participants.

Smart metering, which encompasses the installation of smart meters and the entire AMI, is one of the cornerstones of smart grid deployment. Similarly, you can assist others on the grid by contributing:

a. Customer’s benefits:

With this knowledge and in-home displays, customers will be able to better regulate their energy consumption and expenditures, enabling them to take steps to save energy and/or cut costs. Because to its in-depth knowledge of customer preferences, the utility can give superior, personalized service.

b. Supplier’s benefits:

By omitting the approximated and manual lessons, they will be able to collect virtually real-time readings at a significantly lower price. The mountain of energy usage data will significantly improve the quality of client support. Time-of-use pricing is only one example of the new types of products and services that will emerge, all of which will be aimed to promote customer engagement by providing more personalized options. AMI Organization: The Advanced Metering Infrastructure comprises of multiple components, ranging from the interior of the client's building to the utility itself. Data collection head-ends, meter data management systems, data portals for stakeholder systems, and wide-area networks fall under this category.

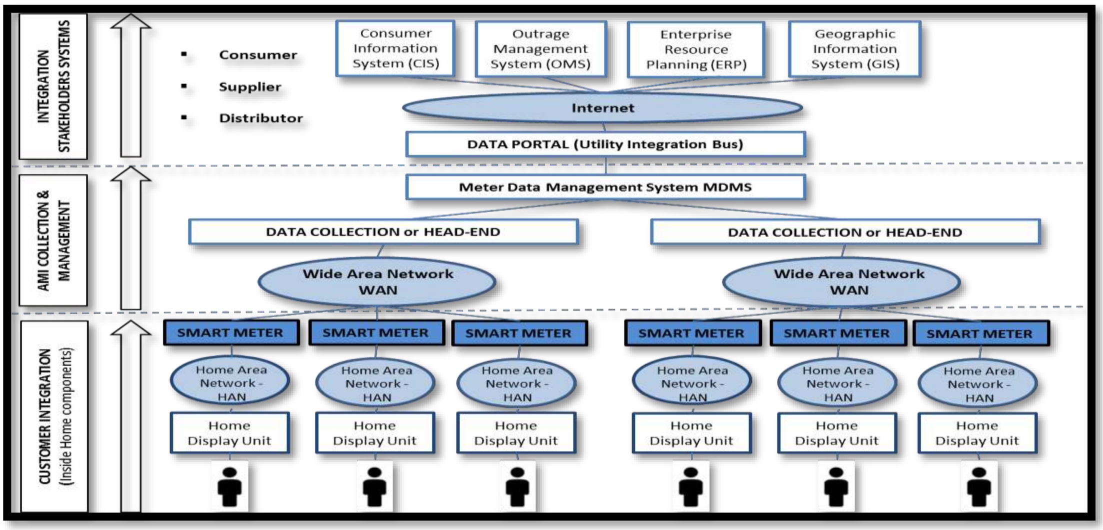


Figure 2.8: Advanced Metering Infrastructure, From Customer's Home to Stakeholder's Services [21].

c. Home displays units:

Many user interfaces provide real-time information on energy use (in kilowatt-hours), expenses, and carbon dioxide emissions. This feature may be restricted to a single manufacturer's equipment, or it may be linked into a personal computer via a website, app on a smartphone, or app on a tablet. Figure 2.9 is an example of a device that can read the data supplied by smart meters.

Figure 2.9: Monitoring Energy Consumption in Different Supports [23].

d. Home or Local Area Networks (Han Or Lan):

Local Area Networks (LANs) connect computers in small areas such as homes, companies, schools, and laboratories. Via gateways, these computers may now exchange information. The Home Area Network (HAN) is a specialized Local Area Network (LAN) that simplifies electronic device and home appliance connectivity.

Figure 2.10 displays the interoperable devices within a HAN, such as smart appliances (refrigerator, washing machine, kitchen, etc.), smart meters, personal computers, tablets, micro-generation from distributed energy resources, and more. While we already use smart devices on a daily basis, the concept of a "smart home" envisions a future in which all appliances are automatically managed by being connected via a number of communication protocols.

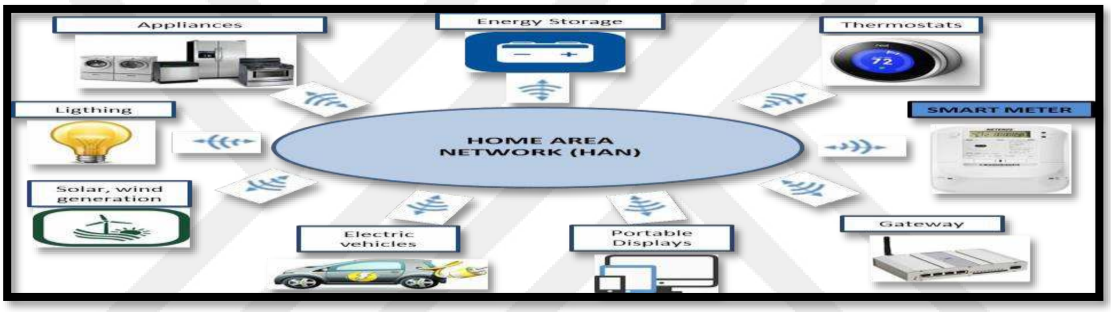


Figure 2.10: Conceptual model diagram of a Home Area Network (HAN) and features [25].

e. Smart Meters

Is essential to the effective running of any AMI. Placed in a residence or company, it supervises and monitors electricity consumption, billing the owner or tenant accordingly. This link allows it to transfer data wirelessly to the utility and the end user.

Figure 2.11: European Smart Meter with Open Smart Grid Protocol [25].

f. Wide Area Networks (Wan)

These networks are capable to cover a broader area, are the medium in which various groups of smart meters from different LANs communicate the data collected to the data concentrator or head end, as illustrated in Figure 2.12.

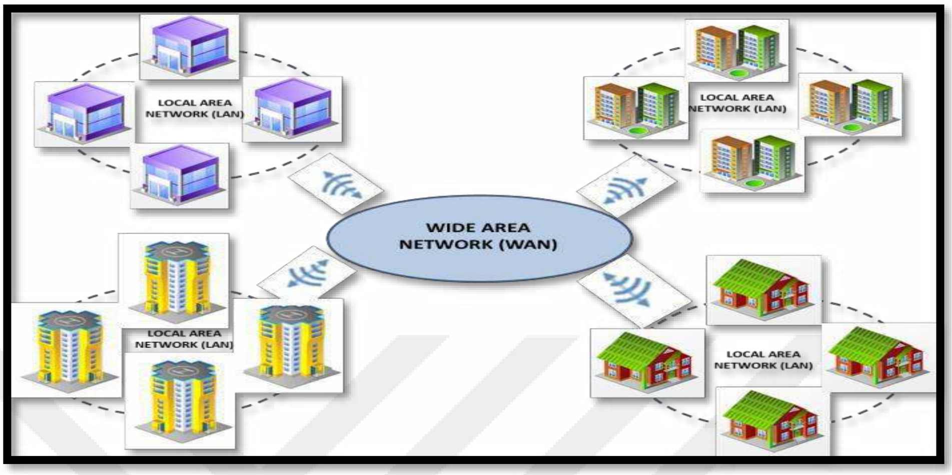


Figure 2.12: Wide Area Network (WAN) Model, Communicating and Aggregating Various Smart Meter Measures from the Home and Local Area Networks [25].

The smart meter can be connected to other AMI components in a variety of ways. The following is a list of possible modes of communication.

In addition to transmitting data through the power lines, these companies also deliver electricity to households and businesses.

Using this technology, data is also transferred through power lines; however, concentrators and repeaters are necessary.

General Packet Radio Service: This service transports data using the mobile operator's WAN for a fee proportional to the quantity of data transferred.

Radio is used to transmit data from intelligent meters.

Wi-Fi (Wireless Fidelity) is a standard for establishing low-cost, high-speed wireless Internet connections.

Centralized data collection and processing:

The head end is responsible for gathering data from an assortment of smart meters. The smart meter and head-end system should use the same protocol for communication.

A protocol is a set of rules for exchanging digitally-formatted data between computers or inside a network. Transmission Control Protocol/Internet Protocol (TCP/IP), IEC, and ANSI are among the protocols that smart meters can utilize.

Meter data management systems (MDMS):

MDMS is a centralized database capable of storing vast quantities of data from smart meters. Due of their interval tracking capabilities, smart meters generate a substantial amount of data. If a single meter takes a reading once each hour or every half-hour, for example, it may produce between 4,000 and 8,000 readings per year.

A data portal is a hub for evaluating and translating raw data into actionable insights that can be shared with relevant parties (customer, distributors, and suppliers). Other information systems that can communicate with the MDMS database include billing customer information systems (CIS), utility websites, outage management systems (OMS), enterprise resource planning (ERP), power quality management and load forecasting systems, geographic information systems (GIS), and transformer load management (see Figure 2.13). (TLM).

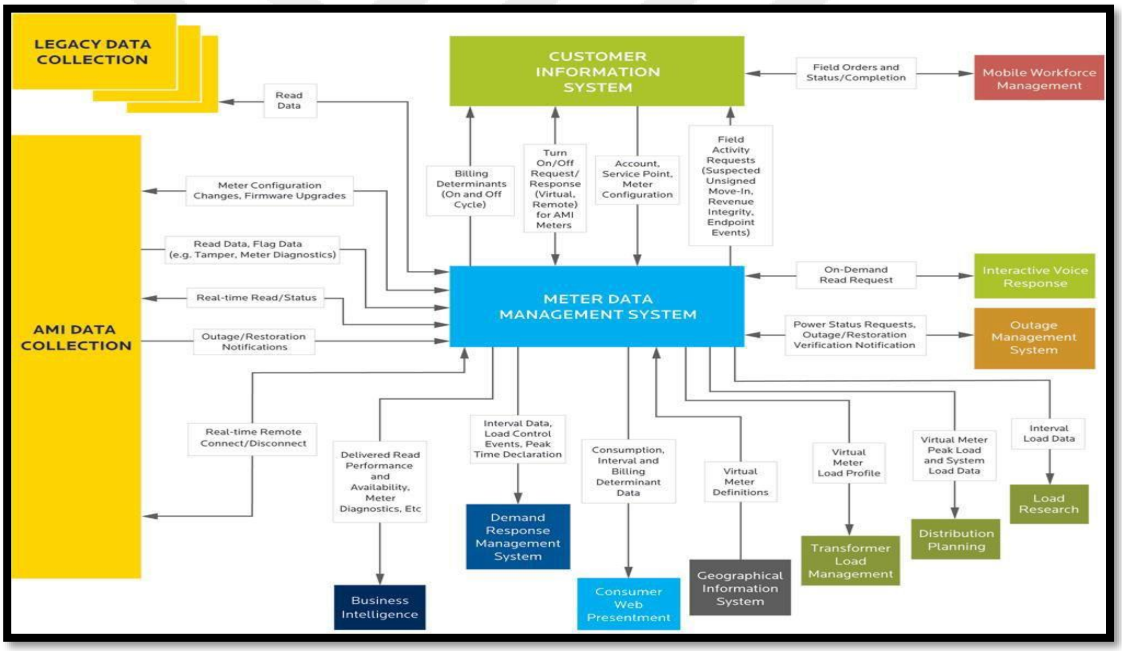


Figure 2.13: Meter Data Management System Inputs/Outputs and Information Systems served [27].

This is the complete process of data gathering, transmission, and analysis within the AMI architecture, which increases the data's value by transforming it into knowledge. The processed data is subsequently transmitted to the consumer or utility, allowing them greater control over their usage or service.

2.2.2 Smart Meters

The smart meter is the foundation of both smart metering and AMI, but its capabilities extend far beyond the monthly readings of energy use that conventional meters offered. And use its interoperability to send this information to the utility and the end user. In addition to electric utilities, water and gas utilities also have smart meters.

The smart meter also performs numerous additional duties. For example, it can inform the user of time-of-use prices, in which costs are greater during peak hours when electricity is scarce in order to stimulate demand response and reduce use during these periods (see Figure 2.14). This information aids the utility in preventing blackouts and enhancing customer service, while also assisting the client in making more informed decisions about when to use electricity.

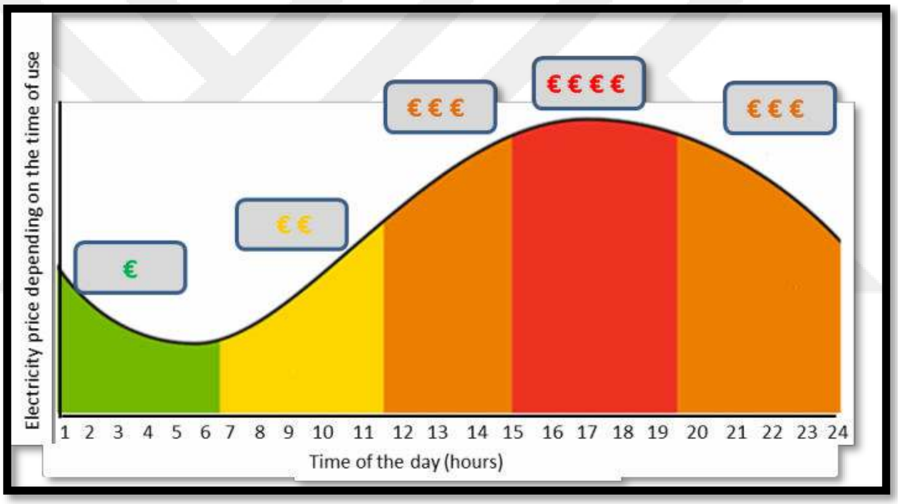


Figure 2.14: Time of Use Tariffs Graphs, Electricity is More Expensive in the Afternoon Where the Peak Demand [23].

enables clients with on-site generation facilities, such as rooftop PV systems, to use their own energy and export any excess to the grid, thereby reducing their energy imports from the grid, as depicted in Figure 2.15. The term "self-consumer" has evolved over the past decade due to the rising cost of energy and the ubiquitous availability of renewable energy sources. Depending on local regulations, net metering may provide a discount on grid electricity or permit the sale of this surplus.

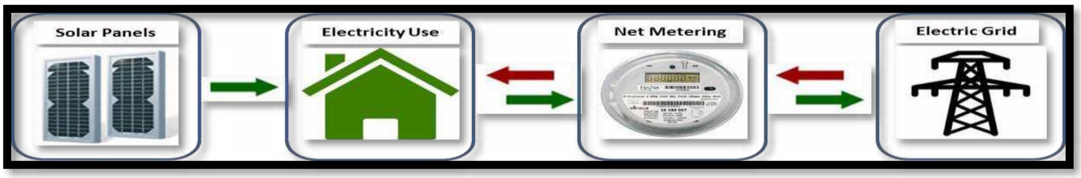


Figure 2.15: Scheme of Net Metering, Two Way Smart Meter Importing and Exporting Electricity with the Grid [29].

As more people acquire and employ Wi-Fi-enabled smart appliances and devices, it will soon be possible to control every aspect of a smart home remotely. This will be possible due to the smart meter's ability to communicate with other smart household devices.

Smart meter installation, Globally, Europe is acquiring digital gas and electricity meters. According to a 2014 survey, there are approximately 200 million digital electricity meters in use worldwide. Even though their installation typically costs between €200 and €250, it has been demonstrated that consumers can save money by utilizing smart meters. Despite predictions that 80% of consumers will have a smart meter by 2020, only 72% will actually have one.

According to "Smart Electric Meters, Advanced Metering Infrastructure, and Meter Communications: Global Market Analysis and Forecasts" [26] by Navigant Research, by 2020, approximately 70% of smart meters will be installed in North America, Europe, and the Asia-Pacific, followed by Latin America with almost 50% and the Middle East & Africa with less than 10%.

In 2014, only eight of North America's fifty states had penetration rates of 75% or higher, while the remaining twenty-eight states had penetration rates below this threshold. Uncertain regulations, decreased electricity rates in some regions, and a lack of enthusiasm on the part of utilities are the primary obstacles to solar energy's expansion.

A collection of laws and regulations govern the European Union.The guidelines include the fundamental requirements for smart meters specified by the European Union. The Europe Energy Directive [12] is the first regulation of smart meters in the region. Table 2.2 (Cost-benefit analyses and the present stage of smart meter adoption) [12] summarizes these characteristics, the essential components that a smart metering system must possess to be beneficial to all stakeholders. As stated previously, the duties and competencies of customers will be crucial. Tracking consumption and providing data to the client frequently (ideally every 15 minutes) can reduce energy demand peaks and costs. The intelligent pricing system is also essential because it can be supported by hourly or half-hourly data collection. For excellent service and client energy management, however, it may be prudent to provide data every 24 hours or more frequently [26-28].

3. METHODOLOGY

The steps required to complete this study and derive the final customer segmentation for estimating power consumption based on utility categorization and clustering are described below.

Obtaining data on the user's hourly power consumption, features of the residence, and occupants, adjusting the data as required for the study, and understanding how the data is displayed are the initial steps prior to analysis. The term "data cleansing" refers to the process of removing inaccurate or superfluous data from a dataset. To complete this step, a visual aid is required.

You may be able to obtain a deeper understanding of the study's findings and verify their veracity through additional reading and data visualization.

Three distinct phases comprise the process: pre-clustering, clustering itself, and post-clustering analysis. Using cluster analysis, the second phase classifies customers into more manageable subsets.

Before making conclusions about the electrical load profiles of users, it is essential to evaluate pertinent studies, select an appropriate input data format, and analyze the input data to determine the proportionate energy consumption per hour for each user.

During the classification phase, the prevalent data clustering techniques for energy use data are evaluated. Weka is used to perform computational clustering using the appropriate hierarchical, naive bayes, and support vector machine algorithms to repeatedly determine the optimal number of clusters and the optimal number of cluster members. Methods for clustering are evaluated based on their aesthetic and statistical outcomes, respectively.

After classifying users, they are redistributed to construct the corresponding consumer subsets. Using visualization and statistical methods, the analyst identifies outliers and manually assigns them to the most suitable group.

The third step is investigating the family and its members. Histograms are used to identify similarities between distinct consumer categories.

3.1 OBJECTIVE OF CLEANING DATA

The first phase of any data analysis is to gather the required data. The data used must also be reliable. It is essential to establish the purpose of the analysis beforehand, as this will determine the data analysis strategy. Using a data preparation service also results in more accurate conclusions.

The required steps are depicted in Figure 3.1:

Figure 3.1: States of the Data When Performing a Data Analysis [25].

When the intent of the data is clear or when we have specific objectives in mind, we can begin to examine the unedited presentation. Although it may be difficult to view and, therefore, comprehend raw data, it is highly unlikely that we will be able to obtain the data in the exact format required to launch the investigation.

This is referred to as "processed" or "tidy" data, and it must be prepared in this manner prior to analysis. With the aid of computational data processing, processing programs are created and executed for this purpose. These scripts contain the fundamental code for data configuration.

After the data have been processed, data cleansing is carried out prior to any analysis. The term "data cleansing" encompasses both preventative and corrective measures taken to rid a dataset of errors such as duplicates, absent data, NAs, and others.

Following are descriptions of the numerous types of data that will be incorporated into this study:

a. Electricity consumption data

b. Households and householder’s information

3.2 ELECRICITY CONSUMOTION DATA

The ensuing investigation heavily relied on data regarding the families' electricity consumption.

Figure 3.2 is a visual depiction of the steps necessary to acquire the pertinent data on electrical usage for the investigation:

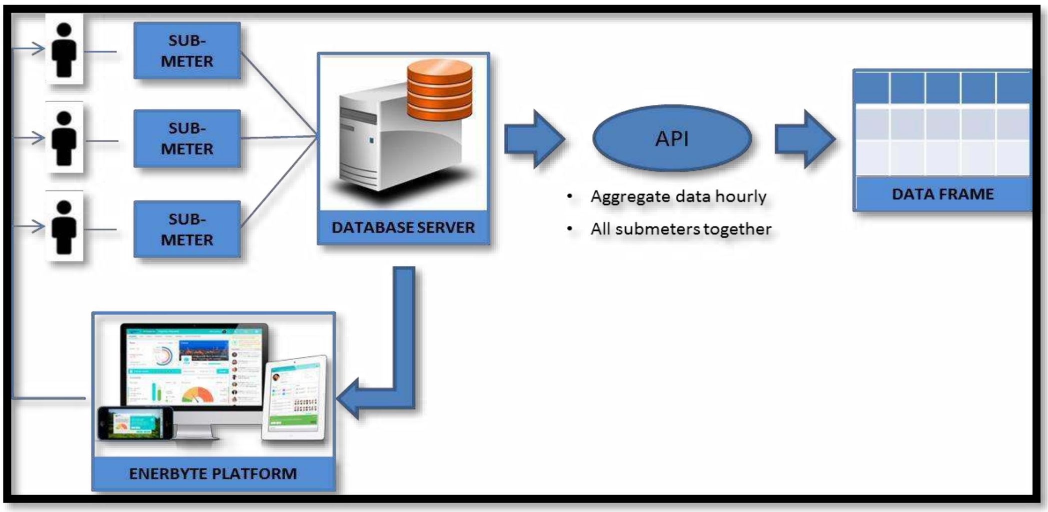


Figure 3.2: Diagram to Explain How the Data Used for the Analysis is Obtained [27].

Submetering devices are used to monitor the quantity of energy consumption in residences and smaller businesses. These devices transfer data via a link, often Wi-Fi in the home, to the manufacturer's database server, where it is kept.

3.3 DATA EXPLORATION AND VISULIZATION

Before beginning consumer segmentation, it is essential to know the data that will be analyzed in order to have a feel for the information and identify any abnormalities or difficult areas that must be addressed. mastering both the raw data and the "Weka software" instructions necessary for its analysis.

Remember these three things when conducting data analysis:

Let's begin with the dataset format, which in the majority of instances must be modified to provide the appropriate visual representation.

The way dates are handled provides a barrier when doing analysis, as time series data necessitate creative methodologies. Especially when distinguishing between weekdays and weekends or between different months.

Finally, pay great attention to the specifics of the statistics on electricity use and investigate the most effective techniques of obtaining the necessary information or conclusions.

Presently, data is collected hourly. For a more in-depth analysis to establish the variables that created the profile shape and peak positions, as well as the characteristics of the home, data with a resolution of 5 or 15 minutes would be necessary.

3.4 OBJECTIVE OF PREDICTION

The objective is to display all the data in a manner that best suits the intended purpose. It is thrilling to provide a service that satisfies all client wants and provides visual representation for each client. The ability to rapidly access the most significant characteristics of each customer's power consumption makes it easy to conduct a full audit of a customer's usage in order to detect anomalies or problems more quickly.

a. One of these objectives is to compare weekday and weekend load profiles to ones that include both weekdays and weekends in equal proportions.
b. Differentiating the load profile for each day of the week will reveal whether the load profiles on different weekdays are equivalent or dissimilar (Friday's load profile may differ from Monday through Thursday's) (Monday through Sunday). Quantitative statistical representation of hourly consumption rates and totals.
c. Determine if there are seasonal and monthly fluctuations in consumption.
d. Identifying the typical monthly load profile; setting discrete time intervals and visualizing usage patterns; Information can be displayed in a variety of ways, including the use of various graph forms and colors (bars, points, lines, boxplots, etc.); interactive graphs can assist users in comparing data.

3.5 ENERGY CONSUMPTION VISUALIZATIONS

This analysis will examine the electricity bill from multiple perspectives. The technique begins with a broad overview of the data and then zeroes in on specific areas by segmenting and categorizing the information for a more targeted analysis.

Both absolute figures of power consumption (kWh) and proportional hourly energy usage in percentages (%) were used to plot the electrical consumption. While the absolute units in kWh are essential for studying each customer separately, a normalized framework is required to compare the shapes of the curves from different users in an equivalent manner. Percentages of hourly energy consumption [21] can be used for this purpose.

As stated previously, the purpose of the analysis should lead the selection of the most suitable parameter. Given the abundance of available tactics for analyzing consumption data, this section provides an overview of a variety of representational techniques for better understanding and servicing customers. This part employs a variety of consumers for the various representations to observe diverse load patterns and assess their accuracy.

There is a notable distinction between weekday and weekend load profiles. Table 3.2 demonstrates that there are no substantial changes between weekday and weekend consumption. This figure, which is pertinent to the current study, can be used to discuss whether the consumer load patterns should be grouped by weekdays alone or by weekdays and weekends as well.

Establishing a rule to assign a consumption value to certain days could enhance the quality of the visualization by eliminating white spaces produced by missing data and enabling more precise comparisons when summarizing the data.

3.6 CLASSIFICSTION OBJECTIVE

The final goal of this research is to provide each subgroup with suggestions and guidelines for predicting its electricity consumption based on the ideal approach for classifying consumers into subgroups based on their consumption load profiles.

The theoretical foundations of this study are comparable to those of the previously described study [28], however instead of dividing customers into predetermined categories, the focus was on establishing the archetypes of typical load profiles.

Our goal is to identify the optimal number of categories for effectively segmenting our clientele. Several classification techniques were employed, and the number of categories was adjusted in many iterations.

3.7 CLUSTERING OVERVIEW

3.7.1 Cluster Analysis Theory

The objective of cluster analysis is to categorize a set of data points into groups based on similarities in one or more defining factors. Thus, there will be a higher degree of resemblance between things within the same group than between those within other groups. This indicates that there are various ways to group identical data. The analyst is responsible for defining the most appropriate similarity/distance metric based on the similarity criteria. The analyst determines the qualities to employ for clustering the data.

Yet, any of the several clustering algorithms and techniques may result in a somewhat different data structure. For the clustering analysis to be successful, the appropriate approach must be chosen. Not every methodology is optimal for every type of problem; therefore, the type of input data to be clustered is a key consideration when choosing a strategy.

To effectively deliver the findings to the audience, the analyst must possess outstanding visualization, analytical, and communication abilities. The steps of cluster analysis are depicted in Figure 3.3.

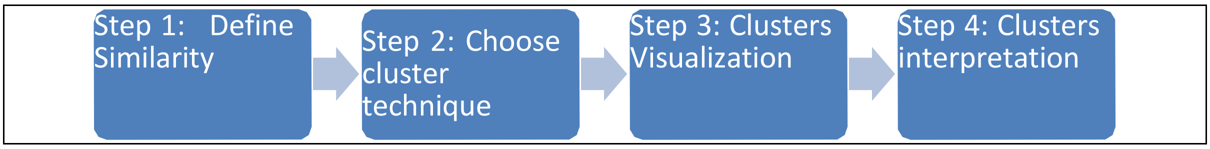


Figure 3.3: Steps to Perform Any Cluster Analysis [21].

Using a technique known as "cluster analysis," which is utilized in statistical analysis and exploratory data mining, allows for the discovery of knowledge, patterns, and information. Several industries, including business and marketing, biology and bioinformatics, computer science, and the digital environment, make extensive use of cluster analysis.

It is essential to emphasize that the data to be clustered in this inquiry are time series data consisting of successive measurements of household and small business power usage recorded over time. This indicates that the electricity bill is itemized per hour.

Using unsupervised learning to group the vectors into map regions, this classification allows for the flattening of high-dimensional data into a 2-dimensional space. The results of employing the Hierarchical technique, an aggregative bottom-up strategy, can be presented in a transparent manner, demonstrating the ideal distance and connection criteria for clustering. The Nave Bayes approach was ultimately chosen since it is the most used clustering technique for vector quantization, despite needing the cluster size to be specified in advance.

3.7.2 Clustering And Classification

At this point it is adequate to differentiate between clustering and classification, see Table 3.1, in order not to misunderstand concepts and define clearly what represents each one:

Table 3.1: Basic Characteristics to Differentiate Clustering and Classification [21].

CLUSTERING	CLASSIFICATION
The similarity characteristics of the data are not known in advance	Exists a training dataset that was used to previously to subset this data in groups
The dataset is divided into clusters or groups. Objects inside the same group have similar properties to each other; and differ from instances of other groups	New data is classified based on the training dataset. Algorithms find the group to which each new object belongs to due to its common characteristics
It is an unsupervised learning, due to the absence of a training dataset able to provide prior knowledge	It is a supervised learning task due to the existence of a training dataset that has labelled or grouped the data previously
The objective is to label or group the observations, according to their similarity	The objective is each new data object into an existing group

3.7.3 Times Series Clustering And Classification

This section is devoted to examining the available clustering and classification strategies to determine the methodology or techniques that will be utilized for this project. It clarifies their fundamental methods, distance parameters, and connecting requirements.

Various clustering and classification algorithms have been employed to classify electrical loads into consistent patterns; a complete literature review on this topic is published in [35], and its results are summarized in Table 3.2.

Table 3.2: Summary of the Clustering and Classification Techniques and Papers, Adapted From [35].

Method	References
Adaptive Vector Quantization (AVQ)	(Tsekouras, et al., 2017)
Random Forest	(Chicco & Sumaili Akilimali, 2010)
Follow-the-leader (FDL)	(Chicco , et al., 2013) (Chicco , et al., 2014) (Chicco , et al., 2015), (Yu , et al., 2015), (Chicco, et al., 2016)
Fuzzy and ARIMA	(Nazarko, et al., 2015) (2015)
Naïve Bayes (NB)	(Chicco , et al., 2015), (Chicco,, 2016), (Tsekouras, et al., 2017)
Decision Tree, Random Forest	(Chicco , et al., 2015), (Chicco, et al., 2016), (Tsekouras, et al., 2017)
K-Linear Regression	(Marques, et al., 2014), (Chicco , et al., 2015), (Chicco, et al., 2016), (Tsekouras, et al., 2017)
Decision Tree	(Verdu , et al., 2016)
Probabilistic Neural Network (PNN)	(Gerbec, et al., 2014), (Gerbec , et al., 2015)
Self-Organizing Map (SOM)	(Marques, et al., 2014), (Verdu , et al., 2016), (Chicco, et al., 2016), (Valero , et al., 2017), (Räsänen, et al., 2010)
Support vector Machine (SVM)	(Chicco & Ilie, 2009)

The table presented a variety of clustering and classification techniques, ranging from agglomerative techniques such as Decision Tree and Random Forest, which do not require a predetermined number of clusters, to partitioning techniques such as Nave Bayes, which do require a predetermined number of classes. According to fuzzy approaches, each instance has a degree of group membership and could potentially belong to multiple groups. Examples of supervised learning include those based on statistics, such as multivariate statistics, and those based on unsupervised learning, such as neural networks and Support vector machines. Support Vector Machine and entropy-based approaches are two other, more contemporary technologies.

This study's objective is to compare the three tactics represented by the three algorithms described above in order to determine which gives the best results in terms of the stated objective of the study.

3.7.4 Distances and Linkages

To determine which observations, belong to a cluster and which classes, it is necessary to develop a measure of similarity or dissimilarity between the data. This is accomplished by employing suitable distance and connecting criteria. A distance or metric separates two observations, whereas linkage divides observations into groups. This article describes Weka software distances and links, as well as some of the more popular ones.

As an example, the "x" and "y" distance vectors are provided. Hence, x and y are the columns.

$$x = (x_1, x_2, \dots, x_n) y = (y_1, y_2, \dots, y_n) \tag{3.1}$$

Euclidean distance: it is the distance of the straight line that between two points, calculated as follows.

$$(x, y) = (y, x) = \sqrt{\sum (y_i - x_i)^2} \tag{3.2}$$

Squared Euclidean distance: is the squared of the Euclidean distance, it is normally used to empathize the difference between objects by placing progressively greater weight on farther objects.

$$d^2(x, y) = d^2(y, x) = \sum (y_i - x_i)^2 \tag{3.3}$$

Manhattan distance: is the distance calculated as the sum of the absolute differences of Cartesian coordinates, also known as rectilinear distance or taxicab distance.

$$(x, y) = \|x - y\|_1 = \sum |x_i - y_i| \tag{3.4}$$

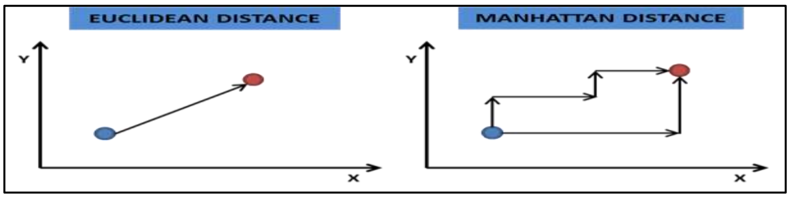


Figure 3.4: Illustration of the Euclidean and Manhattan Distances [45].

3.8 CLASSIFICATION ALGORITHM

Our primary goal is to forecast future electricity use, which may be broadly divided into two categories: The second requirement of an intelligent system is a prediction model that can predict occupancy based on the occupancy data discovered. 1) Possess a categorization model that can reliably determine whether a certain residence is occupied based only on its electricity consumption statistics. In the sections that follow, we will describe the philosophy underlying each of the selected classifiers.

3.8.1 Support Vector Machines

Support vector machines (SVMs) are a prominent technique for applications such as speaker recognition, gene extraction, and facial recognition in the field of supervised machine learning. Kernel approaches can be employed for both linear and nonlinear classification, surpassing neural networks although requiring less training data samples [24]. In addition, SVMs perform effectively when high-dimensional data are involved [31].

In this study, we employed the linear kernel, the radial (or Gaussian) kernel, and the polynomial kernel, which are all SVM models.

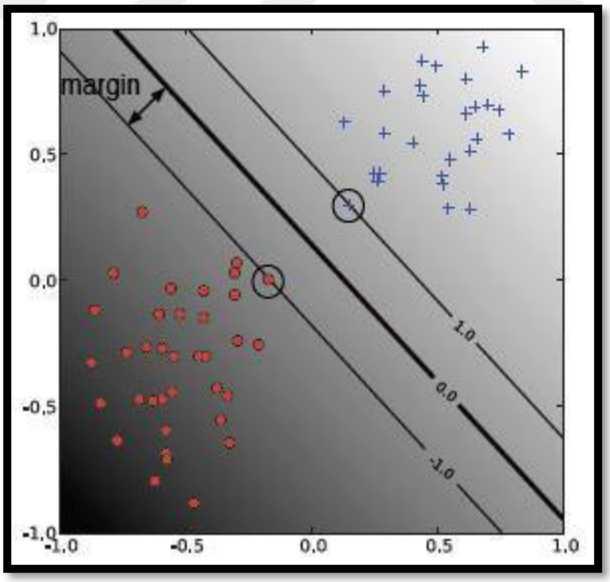


Figure 3.5: Example a Two-Class Linear SVM Classifier [24].

3.8.2 Random Forest

Random forest, a sort of ensemble for machine learning, is frequently used for classification and regression applications. In an ensemble learning approach, predictions are derived by averaging the results of numerous decision trees that have little to no correlation. Using multiple classifiers concurrently reduces model overfitting, resulting in improved classification accuracy. As shown in Figure 3.6, Bootstrap aggregating (also known as bagging) is often employed in ensemble methods. In order to create unique training sets for each classifier, the data is resampled, and new data is introduced.

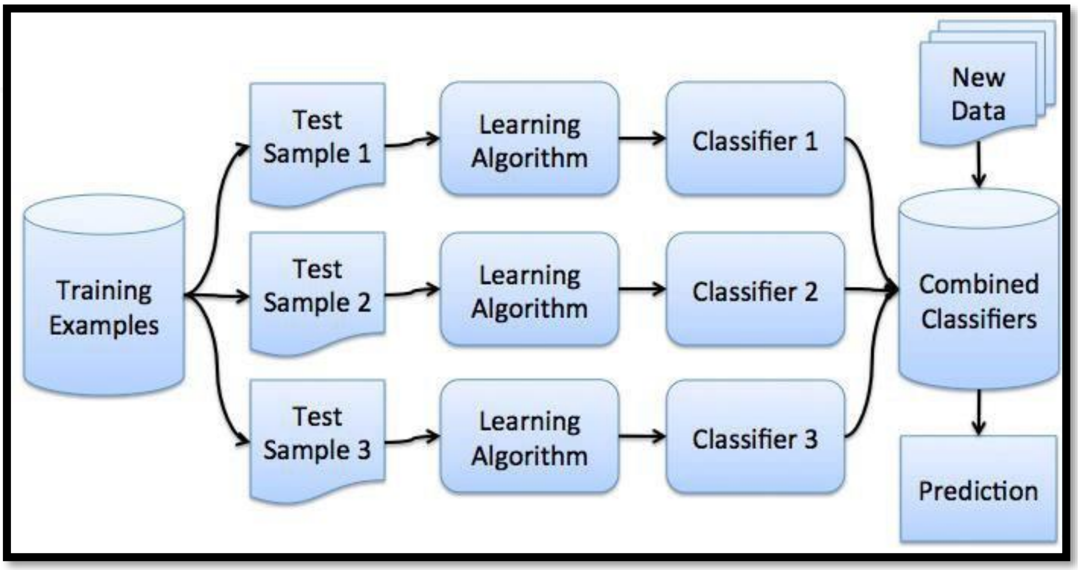


Figure 3.6: Phases of Ensemble Learning Approaches to Solve Classification Problems [31].

3.8.3 Decision Tree

Figure depicts a decision tree consisting of a root node (at1), intermediate nodes (at2, at3, and at4), and leaf nodes (at5) (yes and no outcomes). A decision tree classifies data through recursive partitioning. The branch indicates the result of an evaluation, whereas the root and internal nodes represent the characteristic or characteristic being evaluated. The method divides the tree until it can classify the data accurately without further testing. The outcome of the categorization is kept in the leaf (or terminal) node at the tree's tip [34].

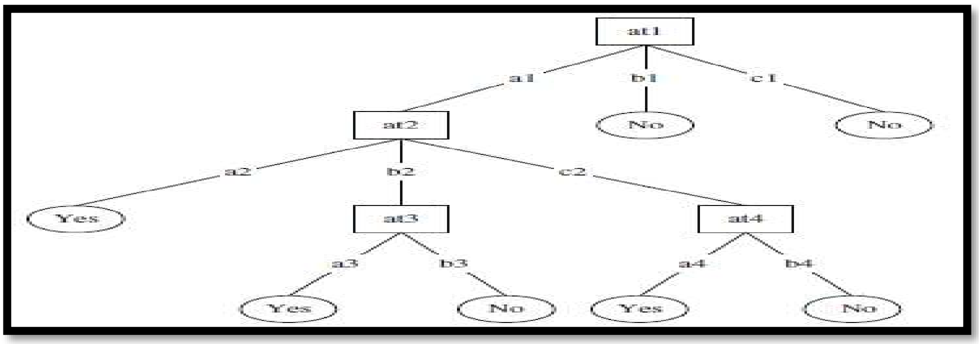


Figure 3.7: Decision Tree Algorithm Structure [34].

3.8.4 Naive Bayes

We build probabilistic classes of habitation using the naive Bayes classification approach, which are subsequently put to the test. The probability that a residence will be involved during a certain weekday and time interval is calculated by dividing the number of involved periods by the total number of existing periods during the specified weekday and time interval. The probability of essence during this time during categorization would be high, for instance, if our collection of characterisation data consisted of weeks of information and the residence was continually assigned to this time. The classifications were then loaded into a naive bayes classifier, which applied the condition to each weekday interval.

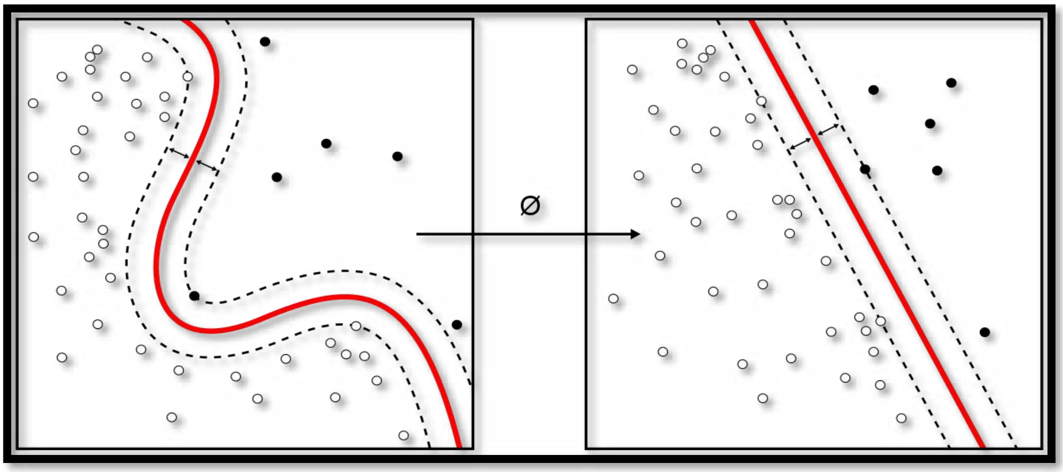


Figure 3.8: Naive Bayes Algorithm on Left with Respect to Support Vector Machine on Right Side For Classification Structure [34].

3.8.5 Hybrid Model

We employ accuracy as the primary criterion for evaluating the efficacy of our order and prediction models, utilizing a decision tree and random forest combination that reflects the degree of correct grouping or expectations. This section provides a further illustration of the use of several measures as a correlative metric. We construct a model and test it in a simulated household to determine how well it predicts occupancy based on energy consumption. Using a traditional model that reliably predicts household occupancy across the board is the primary business problem. To do this, we created a prototype for a single family and then tested it in a huge number of genuine homes.

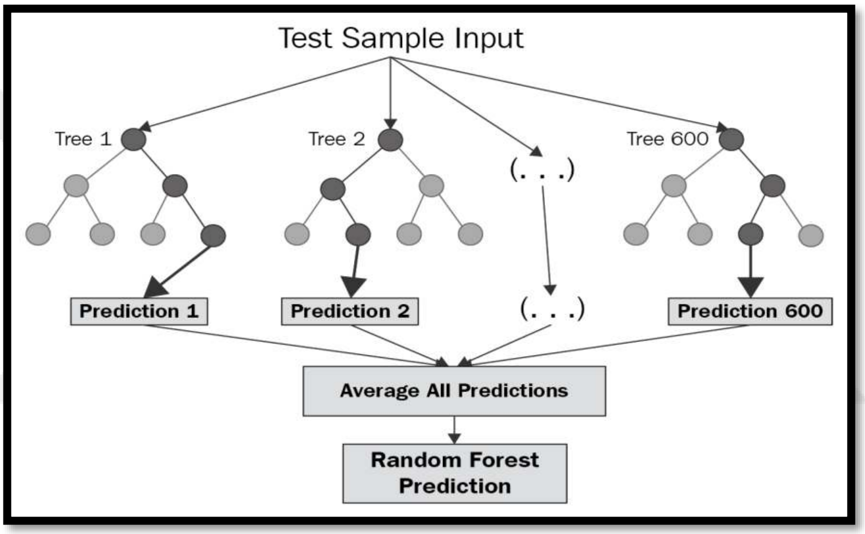


Figure 3.9: Hybrid Model Algorithm Present as Combination of Random Forest and Decision Tree for Classification Structure [4].

.

4. RESULTS

Using a decision tree, a form of tree diagram, the Random Forest approach constructs a taxonomy of categories. Decision Tree and Random Forest can be utilized either agglomerative (bottom-up) or divisively (top-down).

The final step is to classify this prediction based on what will be examined. The analyst behind a schematic forecast may be required to make a difficult choice regarding the number of classes to employ to effectively address the issue at hand. It was decided to only include weekdays since Saturdays and Sundays account for just 30% of overall consumption during a typical week and the load curve over the weekend is significantly different from that of a typical weekday . However, this may change the results.

a. Calculating statistical parameters, such as variance and standard deviation, and determining the minimum, maximum, and median hourly percentages; dividing the percentages into the morning (6:00-11:00h), noon (12:00-17:00h), evening (18:00-23:00h), and night (00:00-05:00h) periods; calculating ratios between the percentage figures and between the periods; etc.
b. Observational trends can be determined by analyzing electricity consumption throughout the week. As consumption habits and weekend activities differ, it is evident that weekdays are superior for defining the load profile. It is vital to identify the load profiles and fit them to the regular load patterns.
c. In addition to analyzing data from all 24 hours of the school day, we also analyze the median, minimum, maximum, and period statistics (night, morning, noon, and evening) (00:00-23:00h).
d. Each piece of data is reviewed both during the week and on weekends.
e. Within the first 24 hours, one class of variables (00:00-23:00h) is utilized; in addition, the median, minimum, maximum, and period numbers (night, morning, noon, evening) are also included.

4.1 FALSE POSITIVE AND FALSE NEGATIVE RATE

Specificity and sensitivity of classification based on classes built for consumption prediction are reflected in the false positive in the findings, which reflects a false positive and may impair the energy efficiency of home heating systems owing to wasted heating.

The false positive rate (FPR) is the ratio of the number of false positives to the total number of available slots.

$$\text{FPR} = \frac{fp}{fp+tn} \tag{4.1}$$

If the classification system has a low percentage of false positives, then it detects absences effectively. Unfortunately, the temperature occasionally drops, which is a false negative and may make the occupants uncomfortable (in a thermostat application). To determine how frequently such errors occur, divide the number of false negatives by the total number of filled slots to obtain the false negative rate (FNR).

$$\text{FPR} = \frac{fn}{fn+tn} \tag{4.2}$$

The classification system's ability to anticipate usage properly is indicative of its low rate of false negatives.

4.2 TRUE POSITIVE AND TURE NEGATIVE RATE

The true positive rate (TPR) is defined as the percentage of positive instances that are correctly classified and is given by equation. It is also known as sensitivity or recall.

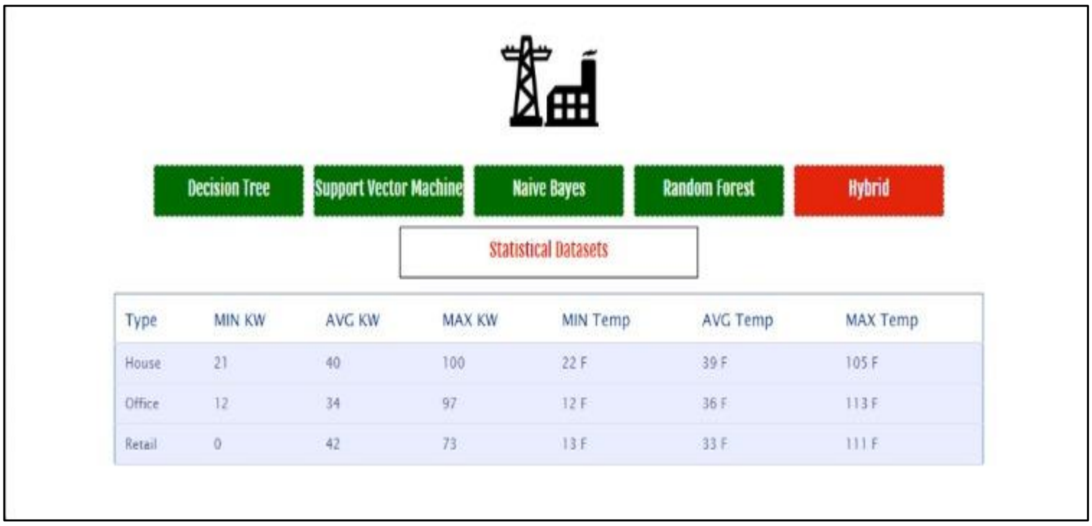
$$\text{TPR} = \frac{tp}{tp+fp} \tag{4.3}$$

The true negative rate (TNR), or specificity, is the proportion of negatives that are correctly identified, and is calculated by equation.

$$\text{FPR} = \frac{tn}{tn+fp} \tag{4.4}$$

4.3 CLASSIFICATION ON STATISTICAL DATASET

In numerous classification schemes, the distance between two classes is defined as the smallest distance between their geometric and statistical centers.

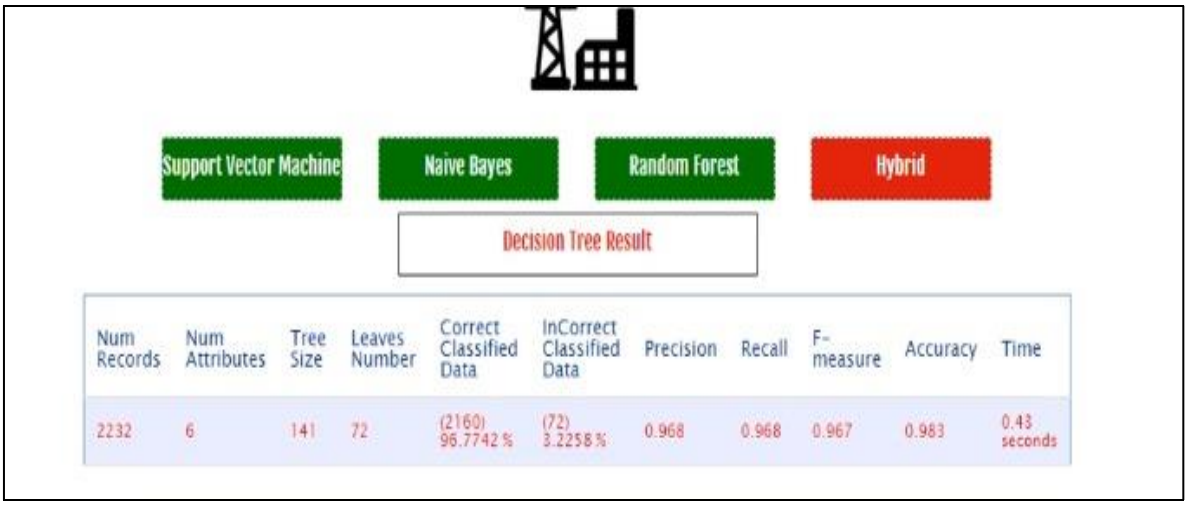


Type	MIN KW	AVG KW	MAX KW	MIN Temp	AVG Temp	MAX Temp
House	21	40	100	22 F	39 F	105 F
Office	12	34	97	12 F	36 F	113 F
Retail	0	42	73	13 F	33 F	111 F

Figure 4.1: The Overall Classification Results Based on Approaches Used on Statistical Dataset.

4.3.1 Decision Tree Prediction

When working with dimensionless data with time facto, such as the number of records, precision, recall, F-measure, and accuracy, the analyst must select the most appropriate input data units for the output results. Depending on the desired conclusion, classification analyses employing decision tree classification and load profile energy consumption segmentation generate diverse sets of output data.



Num Records	Num Attributes	Tree Size	Leaves Number	Correct Classified Data	InCorrect Classified Data	Precision	Recall	F-measure	Accuracy	Time
2232	6	141	72	(2160) 96.7742 %	(72) 3.2258 %	0.968	0.968	0.967	0.983	0.43 seconds

Figure 4.2: The Decision Tree Classification Results Based on Approaches Used.

4.3.2 Random Forest Prediction

Different load profiles have different energy consumption segmentation and classification goals when using random forest classification; using absolute values in result, such as the number of records, precision, recall, F-measure, accuracy, and time for the use of dimensionless data with time trends.

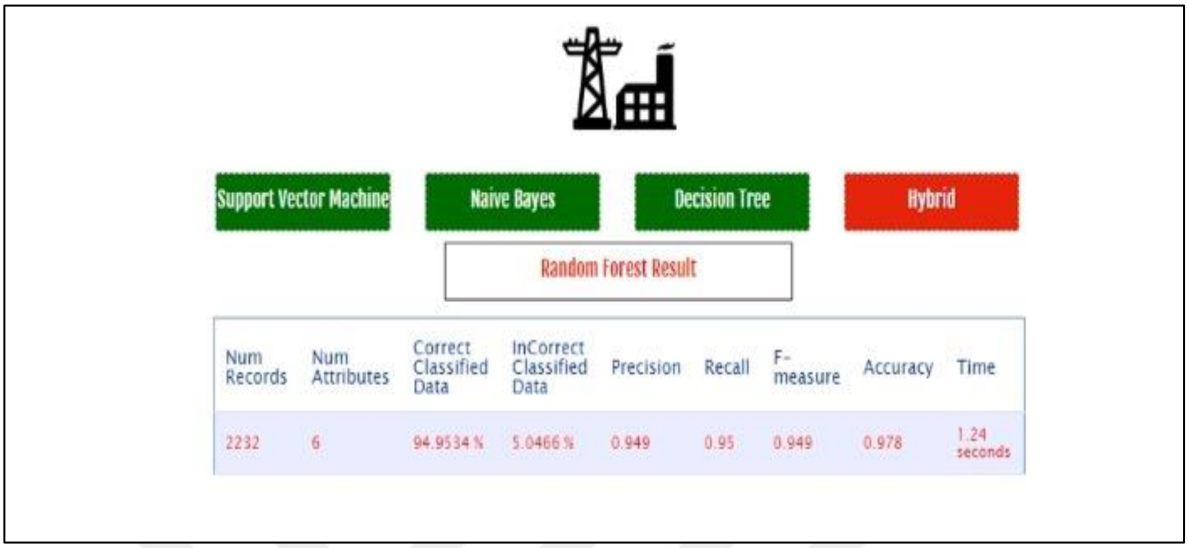


Figure 4.3: The Random Forest Classification Results Based on Approaches Used.

4.3.3 Support Vector Machine Prediction

Depending on the specific analysis goal, different output findings are used to perform the support vector machine classification analysis of the load profile's energy consumption.

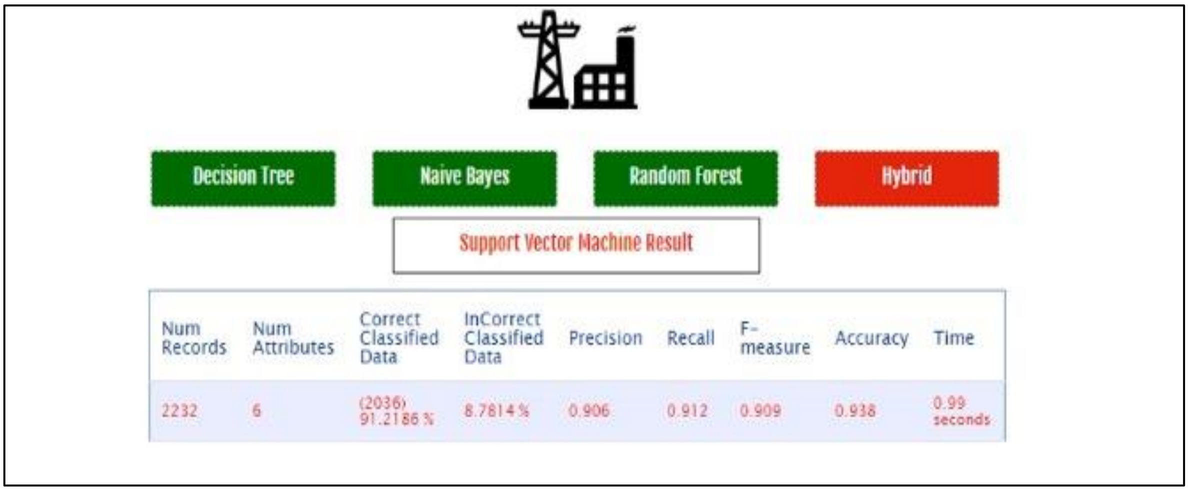


Figure 4.4: The Support Vector Machine Classification Results Based on Approaches Used.

4.3.4 Naive Bayes Prediction

The load profile's energy consumption segmentation and classification using nave bayes classification can serve a variety of reasons; thus, the analyst must select which output data are most valuable for performing the classification study. Absolute numbers are supplied for counts, measures of precision and recall, the F-measure, the accuracy of results, and the time spent on a job for the use of dimensionless data with a time factor for generation and execution being absolute 0.24 seconds and recorded precision of 0.801.

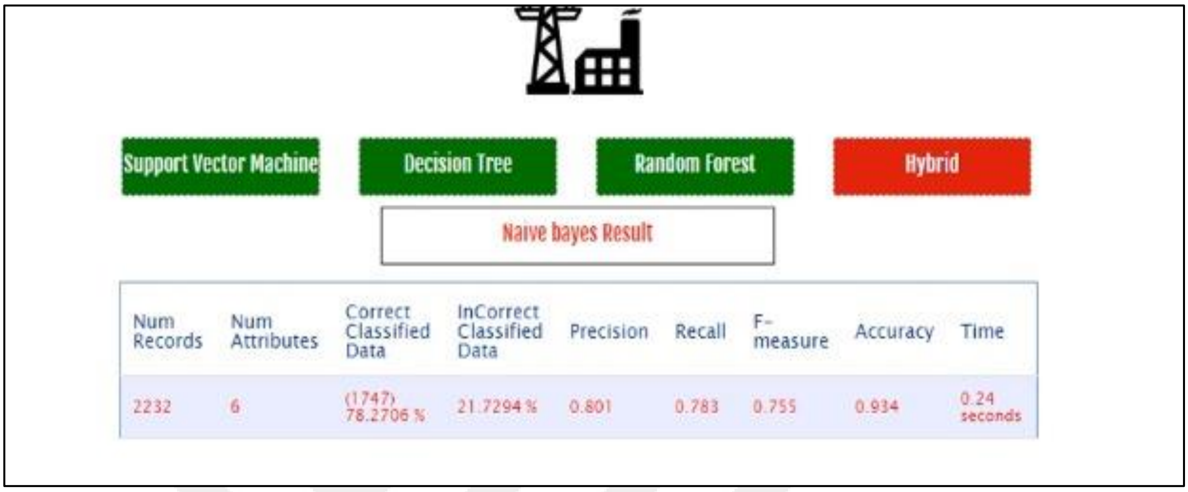


Figure 4.5: The Naive Bayes Classification Results Based on Approaches Used.

4.3.5 Hybrid Model Prediction

The output results used to conduct the classification analysis vary depending on the specific objective of the load profile's energy consumption segmentation and classification using hybrid classification, which deliberately combines decision tree and random forest; the use of absolute values in result, such as the number of records, precision, recall, F-measure, and so on, for the use dimensionless data with time factor for generating and execution is absolute 2.23 seconds with recorded precision of 0.969.

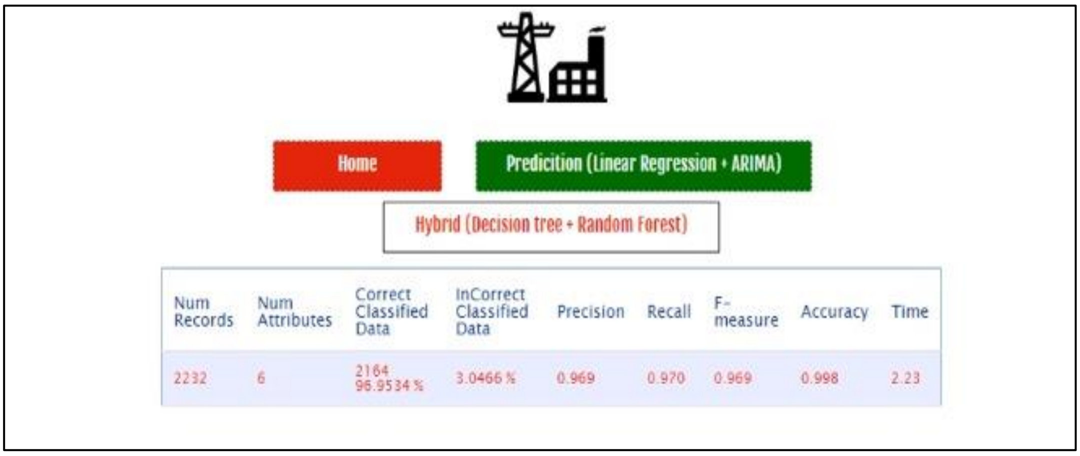
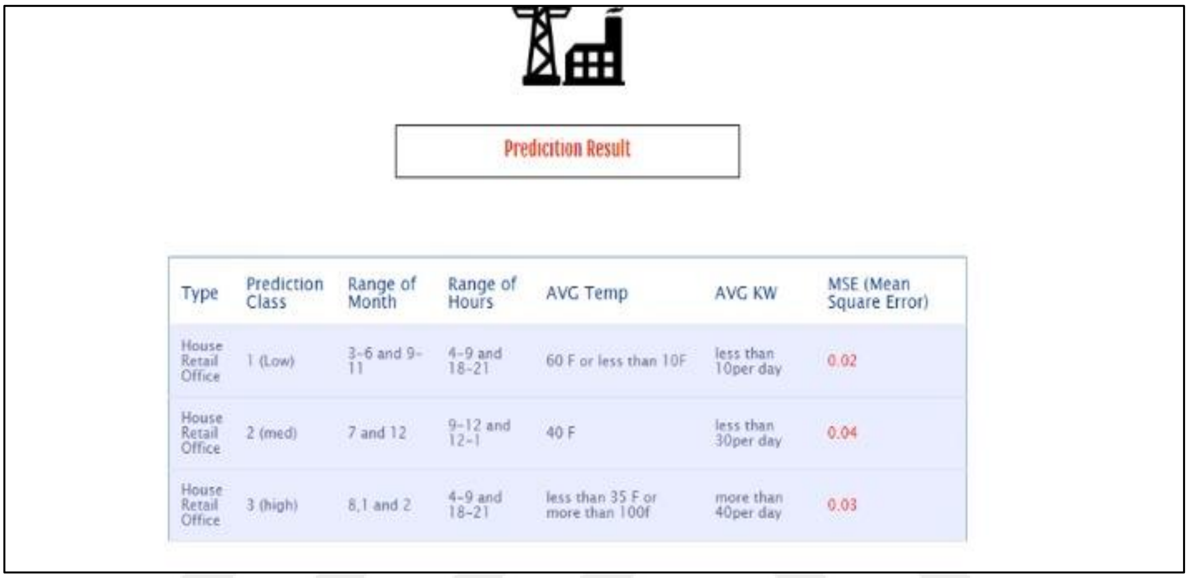


Figure 4.6: The Hybrid Model Classification Results Based on Approaches Used.

4.3.6 Prediction Result

By calculating the AVERAGE KW for the mean square error of each load profile with respect to its class mean load profile, the classification can be validated, and the load profiles that deviate the most from the mean, and thus may need to be reallocated to another prediction class between a range of months based on type, can be identified.



Type	Prediction Class	Range of Month	Range of Hours	AVG Temp	AVG KW	MSE (Mean Square Error)
House Retail Office	1 (Low)	3-6 and 9-11	4-9 and 18-21	60 F or less than 10F	less than 10per day	0.02
House Retail Office	2 (med)	7 and 12	9-12 and 12-1	40 F	less than 30per day	0.04
House Retail Office	3 (high)	8,1 and 2	4-9 and 18-21	less than 35 F or more than 100f	more than 40per day	0.03

Figure 4.7: The Prediction of Result for Classifying the Consumption of Electricity in Terms of Mean Square Error.

5. CONCLUSION

Since home energy consumption is relatively low, energy management systems are not as prevalent in the residential sector as they are in the commercial and industrial sectors. We were able to construct a framework and an intelligent system to predict future electrical energy usage based on these findings. Homeowners may still have access to these services provided the necessary infrastructure is in place, such as when government agencies and electrical companies collaborate to build a community.

This initiative demonstrates the government's early efforts to involve customers in the effort to improve energy efficiency. It aims to educate and direct its users so that they can cut their average energy use by 10%. However, it was discovered that supplying technical consumption data in kWh to households has low impact and influence [38], thus further work is required to translate the technical knowledge into actionable steps that would assist the user in making good energy decisions (like data mining approaches). Scalability is further limited by the necessity for expensive sub-metering devices to collect consumption data. This type of endeavor cannot be expanded without access to data from smart meters installed by utilities.

Thus, the energy business stands to gain significantly from the use of smart meters and big data analytics. Methods of data mining are utilized to increase utility and customer satisfaction by generating large amounts of raw data that must be managed prior to being translated into valuable information. producing data-driven commercial opportunities to suit market needs.

For trustworthy research and policymaking, accurate data on electrical energy usage is essential. Quantitatively and qualitatively, the data used in this investigation were the best available. A substantially larger dataset with thousands of users would have better fulfilled the objectives of the study than the modest sample size that was employed.

Due to difficulties with the electrical consumption data collected by the sub-metering equipment, data cleaning and purification additionally required a pre-clustering stage. Since the sub-metering device and server communicate using the machine learning algorithms we found, consumption data is omitted and consumption is counted as if it had not occurred.

When the possibility of access to hourly electrical energy consumption data becomes a reality, it will provide a suitable framework for the current study's objectives because the number of users could be much higher and the data quality should be nearly perfect, just like the actions taken by the electricity utility to bill their customers using machine learning-based algorithms.

While the primary objective of dividing electrical energy users into groups with comparable load profiles was achieved, the secondary objective of studying the features of these groups added nothing to the project. The common framework utilized to compare and compare fairly the various load profiles, as demonstrated by computing the percentage of energy consumption per hour, served its intended function of classifying users by their load profiles.

We employed an iterative technique based on computer classification computations to arrive at the final clusters (using the Weka program). An outlier was identified, and the analyst manually reassigned it to the appropriate group using visualization and statistical techniques. No matter whatever categorization method was utilized, the outcomes were uniformly comparable (Decision tree, Random forest, Naive Bayes, Support vector machine, or the hybrid model, which combines Random forest and Decision tree). In addition, classes are utilized to aid the depiction of categorization and division possibilities.

5.1 FUTURE WORK

Implement and assess the approach on a bigger and more inclusive data set, such as information collected from smart utility meters. In addition to a more in-depth literature analysis on the treatment and segmentation of smart meter data, more advanced clustering algorithms could be employed to improve specific aspects of the process.

Using weather-normalized electricity consumption data, using more features than just the 24h vector to cluster, eliminating outliers when determining the mean load profile of a user, conducting the same study using the weekend profiles, and applying other and more advanced clustering techniques are methods that can be incorporated into the actual analysis to produce more accurate results.

It was established that while hourly data was sufficient for clustering load profiles, it was insufficient for identifying the origin of the curve's shape and peaks. Thus, it will be essential to use information on equipment and features, as well as statistics on electricity use, within 5 minutes. Owing to the reality that no two consumers are alike, consumer segmentation can assist the utility gain a better understanding of its customers as a whole. Hence, it may be able to serve to various types of customers with specialized offerings. The effectiveness of energy-saving programs will increase, and as a result, energy-reduction recommendations will be improved. Demand response solutions that move energy use away from or shorten peak periods can also be beneficial.

REFERENCES

[1] T. Hove and E. Mungofa, “Towards the Optimal Sizing of Solar-Powered Pump-Pipe-Storage Systems,” Proceedings of the ISES Solar World Congress 2019, 2019.

[2] N. Armaroli and V. Balzani, “Towards an electricity-powered world,” Energy & Environmental Science, vol. 4, no. 9, p. 3193, 2011.

[3] Beckel, L. Sadamori, and S. Santini, “Towards automatic classification of private households using electricity consumption data,” Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, Nov. 2012.

[4] Nguyen Minh, “Solid Oxide Fuel Cell Hybrid System for Distributed Power Generation,” Jul. 2004.

[5] S. Bouzarovski, “Energy poverty in the European Union: landscapes of vulnerability,” Wiley Interdisciplinary Reviews: Energy and Environment, vol. 3, no. 3, pp. 276–289, Aug. 2013.

[6] G. Chicco and I.-S. Ilie, “Support Vector Clustering of Electrical Load Pattern Data,” IEEE Transactions on Power Systems, vol. 24, no. 3, pp. 1619–1628, Aug. 2009.

[7] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Customer Characterization Options for Improving the Tariff Offer,” IEEE Power Engineering Review, vol. 22, no. 11, pp. 60–60, Nov. 2002.

[8] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, “Emergent electricity customer classification,” IEE Proceedings - Generation, Transmission and Distribution, vol. 152, no. 2, p. 164, 2005.

[9] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, “Load Pattern-Based Classification of Electricity Customers,” IEEE Transactions on Power Systems, vol. 19, no. 2, pp. 1232–1239, May 2004.

[10] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” Energy, vol. 42, no. 1, pp. 68–80, Jun. 2012.

[11] K. Carrie Armel, A. Gupta, G. Shrimali, and A. Albert, “Is disaggregation the holy grail of energy efficiency? The case of electricity,” Energy Policy, vol. 52, pp. 213–234, Jan. 2013.

[12] Georgescu, Afshari, and Bornard, “Optimal adaptive predictive control and fault detection of residential building heating systems,” Proceedings of IEEE International Conference on Control and Applications CCA-94, 1994.

[13] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," in Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, 2007.

[14] J. Scott, A. J. Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges and N. Villar, "PreHeat: Controlling Home Heating Using Occupancy Prediction," in Proceedings of the 13th international conference on Ubiquitous computing, 2011.

[15] J. Krumm and A. J. Bernheim Brush, "Learning Time-Based Presence Probabilities," in Pervasive Computing, 2011.

[16] X. Liang, T. Hong e G. Q. Shen, “Occupancy data analytics and prediction: A case study,” em Building and Environment, Elsevier, pp. 179-192,2016.

[17] A. Nadali, E. N. Kakhky, and H. E. Nosratabadi, “Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system,” 2011 3rd International Conference on Electronics Computer Technology, Apr. 2011.

[18] A. Fleury, M. Vacher, and N. Noury, “SVM-Based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results,” IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 2, pp. 274–283, Mar. 2010.

[19] Z.-T. Yu, Y.-X. Qiu, J.-H. Deng, Lu Han, C.-L. Mao, and X.-Y. Meng, “Research on Chinese FAQ Question Answering System in Restricted Domain,” 2007 International Conference on Machine Learning and Cybernetics, 2007.

[20] M. Swain, S. K. Dash, S. Dash and A. Mohapatra, "An Approach for IRIS Plant Classification Using Neural Network," International Journal on Soft Computing, vol. 3, 2012.

[21] T. Durieux, "Exploring the use of artificial neural network based subgrid scale models in a variational multiscale formulation," 2015.

[22] S. K e S. Sasithra, “REVIEW ON CLASSIFICATION BASED ON ARTIFICIAL NEURAL NETWORKS,” International Journal of Ambient Systems and Applications (IJASA), vol. 2, 2014.

[23] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification," Journal of Biomedical Informatics, p. 352–359, 2003.

[24] Y. LeCun, L. Bottou, G. Orr and K.-R. Müller, "Effiicient Back-Propogation," in Neural Networks: Tricks of the Trade, Springer-Verlag, pp. 9-50, 1998.

[25] A. Ben-Hur and J. Weston, "A User’s Guide to Support Vector Machines," in Methods in molecular biology, pp. 223-239, 2010.

[26] J. Ahn, “A stable hyperparameter selection for the Gaussian RBF kernel for discrimination,” Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 3, no. 3, pp. 142–148, May 2010.

[27] E. Goel and E. Abhilasha, "Random Forest: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 1, pp. 251-257, 2007.

[28] Anuradha and G. Gupta, "A Self Explanatory Review of Decision Tree," in IEEE International Conference on Recent Advances and Innovations in Engineering, 2014.

[29] S. B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, p. 261–283, 2011.

[30] Z. Li, “Abnormal Energy Consumption Analysis Based on Big Data Mining Technology,” 2020 Asia Energy and Electrical Engineering Symposium (AEEES), May 2020.

[31] R. R. Rathod and R. D. Garg, “Regional electricity consumption analysis for consumers using data mining techniques and consumer meter reading data,” International Journal of Electrical Power & Energy Systems, vol. 78, pp. 368–374, Jun. 2016.

[32] P. Memari, S. S. Mohammadi, and S. F. Ghaderi, “Data mining model for evaluating and forecasting energy consumption by cloud computing,” 2018 IEEE Electrical Power and Energy Conference (EPEC), Oct. 2018.

[33] S. Singh, A. Yassine, and S. Shirmohammadi, “Incremental mining of frequent power consumption patterns from smart meters big data,” 2016 IEEE Electrical Power and Energy Conference (EPEC), Oct. 2016.

[34] Y. Ni, C. Engström, A. Malyarenko, and F. Wallin, “Investigating the added values of high frequency energy consumption data using data mining techniques,” AIP Conference Proceedings, 2014.

[35] E. Vinagre, L. Gomes, and Z. Vale, “Electrical Energy Consumption Forecast Using External Facility Data,” 2015 IEEE Symposium Series on Computational Intelligence, Dec. 2015.

[36] Z. Feng, J. Huang, W. H. Tang, and M. Shahidehpour, “Data mining for abnormal power consumption pattern detection based on local matrix reconstruction,” International Journal of Electrical Power & Energy Systems, vol. 123, p. 106315, Dec. 2020.

[37] V. Popov, M. Fedosenko, V. Tkachenko, and D. Yatsenko, “Forecasting Consumption of Electrical Energy Using Time Series Comprised of Uncertain Data,” 2019 IEEE 6th International Conference on Energy Smart Systems (ESS), Apr. 2019.

[38] J.-M. Andreoli, “Learning energy consumption profiles from data,” 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Dec. 2014.