Modeling, detection, and prevention of electricity theft for enhanced performance and security of power grid

Soma Shekara Depuru
The University of Toledo
A Dissertation

entitled

Modeling, Detection, and Prevention of Electricity Theft for Enhanced Performance and Security of Power Grid

by

Soma Shekara Sreenadh Reddy Depuru

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Engineering

Dr. Lingfeng Wang, Committee Chair

Dr. Vijay Devabhaktuni, Committee Member

Dr. Mansoor Alam, Committee Member

Dr. Mohamed Hefzy, Committee Member

Dr. Roger King, Committee Member

Dr. Patricia R. Komuniecki, Dean
College of Graduate Studies

The University of Toledo

August 2012
An Abstract of

Modeling, Detection, and Prevention of Electricity Theft for Enhanced Performance and Security of Power Grid

by

Soma Shekara Sreenadh Reddy Depuru

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Engineering

The University of Toledo
August 2012

This dissertation contributes to the development and implementation of novel algorithms for analyzing the electricity consumption patterns of customers and identifying illegal consumers based on irregularities in consumption.

Distribution of electricity involves significant Technical as well as Non-Technical Losses (NTL). Illegal consumption of electricity or electricity theft constitutes a major share of NTL. This dissertation discusses several methods implemented by illegal consumers for stealing electricity and provides relevant literature review. A comprehensive review of the advantages, challenges and technologies involved in the design, development, and deployment of smart meters is presented.

With the advent of advanced metering technologies, real-time energy consumption data will be available at the utilities end, which can be used to detect illegal consumers. This dissertation presents an encoding technique that simplifies the received customer energy consumption readings (patterns) and maps them into corresponding irregularities in consumption. The encoding technique preserves the exclusivity in the energy consumption patterns. The encoding technique saves significant CPU time in the
real-time analysis and classification of customers, in addition to decreasing the memory required to store historical data. Then, this dissertation elucidates operation of intelligent classification techniques on customer energy consumption data to classify genuine and illegal consumers. These classification models are applied on regular energy consumption data as well as the encoded data to compare corresponding classification accuracies and computational overhead.

Further, performance and scope of the proposed algorithms is enhanced in two directions - reducing the overall computation time, and including more real-time parameters using High Performance Computers (HPC). The encoding and classification algorithms are parallelized (in both Task Parallel and Data Parallel approaches). On the other hand, impact of Time-Based Pricing (TBP) and Distributed Generation (DG) on illegal consumers as well as the algorithms used for detection of illegal consumers are analyzed. Economics involved in terms of losses due to illegal consumption of electricity is also explained.
To my family and friends
Acknowledgements

I would like to express my deepest gratitude to Dr. Lingfeng Wang, Adviser and Dr. Vijay Devabhaktuni, Co-Adviser for their support and guidance throughout my PhD program. They have always been with me and taught me how to be productive and successful. I wholeheartedly appreciate the freedom they gave me to do my research. I am thankful for their insightful coaching, tremendous support and encouragement.

I would like to express my sincere thanks to Dr. Mansoor Alam for giving me several career and personal advices, which will help me throughout my life. Thank you for being a mentor and supporting me in every possible way.

I thank Dr. Mohamed Samir Hefzy and Dr. Roger King for kindly agreeing to serve on my Dissertation committee.

It has been a great pleasure working with the faculty, staff, and friends at the University of Toledo, during my journey through the Doctoral program. I appreciate the Electrical Engineering and Computer Science (EECS) department and College of Engineering for funding my Doctoral program.

Finally, I would like to thank my family and friends for their encouragement and invaluable support throughout my education. I thank my family for believing me.
# Contents

Abstract iii  
Acknowledgements vi  
Contents vii  
List of Tables x  
List of Figures xi  
List of Abbreviations xvi  

## 1 Introduction  
1.1 Motivation..................................................................................................................1  
1.2 Dissertation Objectives ............................................................................................4  
1.3 Organization of Dissertation ....................................................................................5  

## 2 Measures and Methods for Controlling Electricity Theft  
2.1 Methods of Stealing Electricity ................................................................................6  
2.2 Factors That Influence Illegal Consumers .................................................................9  
2.3 Literature Review.......................................................................................................10  
2.4 Summary....................................................................................................................14  

## 3 Smart Meters for the Power Grid – Challenges, Issues, Advantages and Status  
15
## List of Tables

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Table Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Types of agricultural customers based on the ranges of energy consumption.</td>
<td>26</td>
</tr>
<tr>
<td>4.2</td>
<td>Types of commercial customers based on the ranges of energy consumption.</td>
<td>26</td>
</tr>
<tr>
<td>4.3</td>
<td>Types of residential customers based on the ranges of energy consumption.</td>
<td>26</td>
</tr>
<tr>
<td>4.4</td>
<td>Time zones and corresponding inputs.</td>
<td>34</td>
</tr>
<tr>
<td>4.5</td>
<td>Instantaneous energy consumption readings received for one customer in a single day.</td>
<td>41</td>
</tr>
<tr>
<td>4.6</td>
<td>Encoded readings or pattern of irregularities generated by the encoding process.</td>
<td>42</td>
</tr>
<tr>
<td>6.1</td>
<td>Stages involved in execution of the entire algorithm.</td>
<td>85</td>
</tr>
<tr>
<td>6.2</td>
<td>CPU time for serial execution of the algorithm.</td>
<td>86</td>
</tr>
<tr>
<td>6.3</td>
<td>CPU time for Task parallel execution of the algorithm.</td>
<td>86</td>
</tr>
<tr>
<td>6.4</td>
<td>CPU time for Data parallel execution of the algorithm.</td>
<td>87</td>
</tr>
<tr>
<td>6.5</td>
<td>CPU time for combined parallel execution of the algorithm.</td>
<td>87</td>
</tr>
<tr>
<td>6.6</td>
<td>Average times for ten attempts of execution.</td>
<td>88</td>
</tr>
<tr>
<td>6.7</td>
<td>Average Speedups achieved.</td>
<td>88</td>
</tr>
</tbody>
</table>
List of Figures

1-1: Overview of T&D losses (for an entire year) in different countries [11]..................... 3
2-1: Tapping electricity directly from a distribution feeder - bypassing the meter [16]..... 7
2-2: Technique used by illegal consumers to regulate the supply voltage and manipulate the energy meter reading [17]................................................................. 9
3-1: Framework of communication and electricity networks in a smart grid environment [41]..................................................................................................................... 18
4-1: Instantaneous energy consumption pattern of a residential customer on a weekend for three seasons............................................................................................................. 28
4-2: Instantaneous energy consumption pattern of a different residential customer on a weekday for three seasons. ......................................................................................... 28
4-3: Overall load on feeder for a neighborhood with total losses of 30.9 % on a summer day.................................................................................................................................. 29
4-4: Approximate energy consumption pattern of a small customer on a weekday in summer.................................................................................................................................. 30
4-5: Energy meter readings collected from a consumer’s smart meter, projected for three modes of consumption - genuine consumption, partial and complete pilfering ...... 30
4-6: First stage of the encoding process – this stage simplifies the energy consumption data.................................................................................................................................. 32
4-7: Modifications done as a part of second stage of the encoding process – this stage
modifies the result from first stage of encoding. ................................................................. 37

4-8: Instantaneous energy consumption of one customer over an entire day – an example extracted from the developed data. ................................................................. 38

4-9: Instantaneous energy consumption readings of 440 customers recorded at a specific time. Large variations represent wide range of customers on grid. ................... 38

4-10: Calculated average energy consumption values for 440 customers at a specific time, as a part of the encoding process. ................................................................. 39

4-11: MFs calculated during the encoding process for 100 customers at a specific time. 39

4-12: MFs calculated during the encoding process for a different set of 100 customers at a specific time. ............................................................................................... 40

4-13: Irregularities of 200 customers at a specific time obtained as a result of the encoding process. ............................................................................................... 40

4-14: Irregularities of 20 customers at a specific time obtained as a result of the encoding process. ............................................................................................... 41

4-15: Comparison between a customer’s instantaneous energy readings received and CPC calculated for a single day................................................................. 42

4-16: Pattern of irregularities obtained as a result of complete encoding of energy readings of one customer in one day................................................................. 43

5-1: Operational flowchart of the proposed SVM based classification model. ........... 48

5-2: Graph that illustrates the output classes predicted by the SVM classification model with respect to customer ID................................................................. 51

5-3: Classes predicted by the SVM classification model plotted against actual classes (of consumers) and customer IDs. ................................................................. 51

xii
5-4: Energy consumption readings of one customer mapped against another random customer. .............................................................................................................................................. 52

5-5: Detailed view of the lower cluster of data points in the Figure 5-4. .......................... 52

5-6: Probability that customers are classified into three output classes – represented using different colored lines. .............................................................................................................................................. 53

5-7: CPU times for different attempts of classification by the SVM classification model for normal and encoded energy consumption data. .............................................................................................................................................. 55

5-8: CPU times for different attempts of classification by the SVM classification model for encoded energy consumption data. .............................................................................................................................................. 56

5-9: CPU times for different attempts of classification by the SVM classification model for normal energy consumption data. .............................................................................................................................................. 56

5-10: Classification accuracies of the SVM based classification model for different attempts of classification (different sets of model parameters). ......................................................... 57

5-11: CPU times for different attempts of classification by the SVM classification model for different attempts of classification (sets of model parameters). ......................................................... 57

5-12: Classes predicted by the SVM classification model mapped over original classes in the testing data. .............................................................................................................................................. 58

5-13: Classification accuracies of the SVM classification model plotted against γ values (model parameter) for a probability estimate of one. ................................................................. 59

5-14: Classification accuracies of the SVM classification model plotted against γ values (model parameter) for a probability estimate of zero. ................................................................. 60

5-15: Structure of the chromosome used in the developed NN model. ................................. 62

5-16: Procedure for generating the data required to train the NN model and estimate SVM
model parameters using the NN model.............................................................. 62
5-17: GA fitness graph for the developed NN Model.............................................. 63
5-18: Synopsis of the rules used to develop the rule engine based classification algorithm.
....................................................................................................................... 65
5-19: Sum of individual irregularities in energy consumption of 450 customers’ over a
single day. .................................................................................................................. 66
5-20: Classes classified by the rule engine classification algorithm plotted against actual
classes of consumers. ............................................................................................... 66
5-21 (a): Multilevel classification algorithm for identifying illegal consumers. .......... 68
5-21 (b): Actions proposed to be taken on illegal consumers................................. 68
5-22: Neural network structure of the NNPR based classification model...................... 70
5-23: ROC curves generated by the NNPR classification model............................... 70
5-24: Classes predicted by the NNPR based classification model plotted against the actual
classes of the testing data. ....................................................................................... 71
6-1: Classification accuracy of the SVM based classification model for different attempts
of classification i.e. for different sets of model parameters...................................... 79
6-2: Sum of individual irregularities in energy consumption of 450 customers’ over a
single day. .................................................................................................................. 80
6-3: Classes predicted by the rule based classification algorithm plotted against actual
consumer classes........................................................................................................ 81
6-4: Task Parallel implementation of the entire multi-level classification algorithm. .... 83
6-5: Data Parallel implementation of the entire multi-level classification algorithm...... 84
6-6: Alternative process for parallelizing the proposed multi-level classification
Algorithm. ........................................................................................................................................ 84

7-1: Day Ahead Pricing (in $/kWh) used for billing RTP customers [103]. ...................... 93

7-2: Algorithm to be used for compensating the impact of DG on electricity consumption
to detect illegal consumers.............................................................................................................. 97

7-3: Overall load on the grid in a service region within the PJM RTO [107]. ............ 100

7-4: Total money to be collected by the utility from all customers on the grid in a single
day.................................................................................................................................................. 100

7-5: Money lost by the utility in a single day due to illegal consumption of electricity. 101
List of Abbreviations

AI .................. Artificial Intelligence
AMI .................. Advanced Metering Infrastructure
AOC .................. Average Original Consumption
ATC .................. Average Total Consumption
BPL .................. Broadband over Power Line
CPC .................. Calculated Probable Consumption
CPP .................. Critical Peak Pricing
DG .................. Distributed Generation
DSM .................. Demand Side Management
EDC .................. Electric Distribution Companies
ELM .................. Extreme Learning Machine
FPGA .................. Field-Programmable Gate Array
GA .................. Genetic Algorithm
GIS .................. Geographic Information System
GPRS .................. General Packet Radio Service
GPU .................. Graphic Processing Units
HAN .................. Home Area Network
HPC .................. High Performance Computers
IGBT .................. Insulated Gate Bipolar Transistor
KESC .................. Karachi Electric Supply Company
MAN .................. Metropolitan Area Network
MF .................. Multiplication Factor
NN .................. Neural Network
NNPR .................. Neural Network Pattern Recognition
NTL .................. Non-technical Losses
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2P</td>
<td>Peer-to-Peer</td>
</tr>
<tr>
<td>PLC</td>
<td>Power Line Communication</td>
</tr>
<tr>
<td>RTP</td>
<td>Real-Time Pricing</td>
</tr>
<tr>
<td>SCTP</td>
<td>Stream Control Transmission Protocol</td>
</tr>
<tr>
<td>SIP</td>
<td>Session Initiation Protocol</td>
</tr>
<tr>
<td>SMP</td>
<td>Systematic Multiprocessing</td>
</tr>
<tr>
<td>SMTP</td>
<td>Simple Mail Transfer Protocol</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>T&amp;D</td>
<td>Transmission and Distribution</td>
</tr>
<tr>
<td>TBP</td>
<td>Time-Based Pricing</td>
</tr>
<tr>
<td>TOU</td>
<td>Time-Of-Use</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>VEMS</td>
<td>Vigilant Energy Metering System</td>
</tr>
<tr>
<td>VoIP</td>
<td>Voice over Internet Protocol</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide Area Network</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The Generation, Transmission and Distribution (T&D) of electricity involve huge operational losses. The magnitude of these losses is rising at an alarming rate in several countries. In order to identify illegal consumers of electricity in the view of enhancing the economy of utilities, efficiency and security of the grid, a new method of analyzing electricity consumption patterns of customers and identifying illegal consumers is proposed and realized.

In this chapter, Section 1.1 and Section 1.2 discuss the motivation and objectives behind the development of a technique for detection of illegal consumers in a power grid, followed by Section 1.3, which presents the organization of the dissertation.

1.1 Motivation

Losses that occur during generation can be technically defined, but T&D losses cannot be quantified completely from the sending-end information. Distribution losses in several countries have been reported to be over 30%. Substantial quantity of losses proves the involvement of Non-Technical Losses (NTL) in distribution. Total losses during T&D can be evaluated from the information like total load and the total energy
billed, using established standards and formulae [1]. In general, NTL are caused by the factors external to the power system. Electricity theft constitutes a major chunk of the NTL. Major forms of electricity theft include bypassing (illegal tapping of electricity from the feeder), tampering the energy meter, and physical methods to evade payment [2], [3]. Electricity theft can be defined as, using electricity from the utility without a contract or valid obligation to alter its measurement [3]. Worldwide T&D losses are more than the total installed generation capacity of countries such as Germany, the UK, or France. It is estimated that utilities (worldwide) lose more than $25 Billion every year due to illegal consumption of electricity. For example, utilities in India lose around $4.5 billion every year due to electricity theft [4] and a recovery of about 10% NTL can conserve about 83,000 GWh of electric power annually [5]. In Pennsylvania, PPL, a utility reports an increase in electricity theft by 16% compared to 2008. It has also been identified that the illegal consumption of electricity by local business sector is increasing [6]. Electricity worth approximately $14 million was pilfered in 2010 in the Houston area [7]. In one year, Tampa Electric Company has seen a 20% rise in electricity theft, whereas, Progress Energy has seen an increase between 15 and 20% [8]. Cost of nationwide electricity theft in USA is about $1–6 Billion every year [8], [9]. In Canada, BC Hydro reports that the electricity theft costs $100 million every year [10]. Figure 1-1 shows overall T&D losses in several countries. It is evident that Billions of kWh of energy is being pilfered every year in several countries.

Total losses incurred by utilities due to electricity theft are huge. As the impact of these losses is huge, it is essential to force the implementation of a mechanism that reduces NTL. Quality of the power generated, transmitted, and distributed, influences the
power system components, as well as customer appliances. Illegal consumption of electricity makes the estimation of overall load in real time very difficult. However, parameters involved in analyzing electricity theft include political, economic, criminal, and managerial. In addition, priorities in investment on implementation of new measures might also be prone to corruption.

Figure 1-1: Overview of T&D losses (for an entire year) in different countries [11].

Design of future electricity markets is aimed at providing consumers with highly reliable, flexible, readily accessible, and cost-effective energy services by exploiting advantages of both large centralized generators, as well as small distributed power generation devices [12]. Algorithms proposed in this dissertation could lead to a motivated and strategic progression towards achieving objectives and features of smart grid. Implementation of the smart grid enhances monitoring, control, and optimization of several power quality parameters. In addition, smart meters provide an opportunity to better monitor and control household energy consumption by both customer and utility. After implementation of smart grid, tampering with the meter may be the predominant
way of pilfering. Such tampering methods are yet to be known; yet, may include physical tampering, installation of bugs or software that manipulates energy consumption. Recent reports show that a major utility in Puerto-Rico may have lost up to $400 million in revenue solely due to manipulation of smart meters [13] [14].

1.2 Dissertation Objectives

Detection of illegal consumers is an extremely challenging problem in today’s power engineering and utility’s everyday operations. This dissertation presents a generalized algorithm that uses customer energy consumption patterns to detect illegal consumers in a smart grid environment. To realize this solution, initially, this dissertation conducts an extensive survey on the methods implemented in pilfering electricity and technologies involved in smart energy meters. Then, an extensive survey on the smart meter and communication technologies are carried out explaining the features of smart grid. In general, utilities collect real-time energy consumption information from its customers several times every day. However, owing to the unavailability of that data, a dataset with near real-time energy consumption patterns has been developed in this work. Then, an encoding algorithm is proposed and implemented, which maps instantaneous customer energy consumption patterns into irregularities in consumption, while preserving the uniqueness in patterns of different customers. Then, intelligent classification algorithms are developed and implemented to identify illegal consumers. The proposed algorithms are then modified to be implemented on High Performance Computers (HPCs) for faster analysis and identification. Finally, an analysis on the
impact of Real Time Pricing (RTP) and Distributed Generation (DG) sources on illegal consumption of electricity is presented.

1.3 Organization of Dissertation

The rest of the dissertation has been organized in the following way: Chapter II discusses several methods implemented for pilfering electricity and also provides relevant literature. In addition, chapter II analyses several issues and setbacks in implementing measures for controlling illegal consumption. Chapter III provides an extensive review on smart meter technologies. Chapter IV illustrates the importance of detecting illegal consumers by analyzing customer energy consumption patterns and presents a novel encoding scheme that enhances the process of detecting illegal consumers. Chapter V presents intelligent data classification algorithms used for detection of illegal consumers and discusses the obtained results. Chapter VI illustrates the procedure implemented in modifying the encoding and classification algorithms for exploiting the advantages of HPC. Chapter VII presents an analysis on the impact of TBP and DG on illegal consumers and complications involved in calculating the overall losses. Finally, the dissertation is concluded in Chapter VIII.
Chapter 2

Measures and Methods for Controlling Electricity Theft

In general, electricity consumers may be generalized as genuine customers, partial illegal consumers, and illegal consumers. This chapter presents several simple and sophisticated methods used in pilfering electricity, discusses factors that influence illegal consumers to steal electricity, and reviews technical and non-technical measures proposed or implemented in the detection of electricity theft.

2.1 Methods of Stealing Electricity

The most common and simplest way of pilfering electricity is tapping energy directly from an overhead distribution feeder as shown in Figure 2-1. The next most prominent method of electricity theft is the manipulation of energy meters that are used for recording and billing industrial, commercial and household energy consumption. Though there are many techniques for tampering with such meters, some of these may include:
Exposing meters to strong magnetic fields to wipe out the memory.

Inserting a film or depositing high viscous fluid to disturb the rotation of disc.

Implementation of sophisticated technologies like remote sensing devices.

Tampering the crystal frequency of integrated circuits.

Creating a link between the breaking control wires in an energy meter would divert the current reading in the meter reflecting zero reading at all times.

In the case of electronic meters, Radio Frequency (RF) devices are mounted to affect the accuracy of the meter.

A shunt is installed between the incoming and outgoing meter terminals.

Inter-changing the incoming and outgoing terminals of the meter.

Damaging the pressure coil of the meter.

Resetting the meter reading.

Introducing unwanted harmonics [15].

Exposing the meter to mechanical shock.

Voltage is regulated from the meter terminals, making it read lesser quantity then the original consumption.

Figure 2-1: Tapping electricity directly from a distribution feeder - bypassing the meter [16].
Other engineered methods of tampering with the meter without damaging its terminals are illustrated below. Two-watt hour meters (employed for measuring the energy consumed by large loads with three phase electric supply) are tampered according to the following process: Damage the terminal seal; connect one of the load terminals to the ground; and open the ground wire from the energy meter. In the case of three phase meters, phases are shifted to lower the power consumption reading by the energy meter. Another popular way of lowering the energy meter reading without directly tampering with the meter is shown in Figure 2-2. Here, supply voltage is regulated to manipulate the meter reading. Illegal consumers accomplish this by using one of the three phases; disconnect neutral from the distribution feeder, and using a separate neutral for the return path. Therefore, the energy meter assumes that the voltage between the connected phase and this new neutral is zero, implying that the total energy consumed is zero. Another way of stealing electricity is by isolating neutral and disturbing the electronic reference point by physically damaging the meter. The voltage to be read by energy meter can then be manipulated by controlling the neutral.

In general, illegal consumption of electricity will be predominant only at desired hours of the day - when the customer’s demand is high i.e. using legal electricity for small household loads and illegally tapped electricity for heavy loads. This kind of theft (partial illegal consumption) is very difficult to measure, as the energy consumption pattern is uneven over a period of time. In addition, corrupt employees are often responsible for billing irregularities; they record an amount of consumption that is lower than the original consumption. On the other hand, improper calibration and illegal de-
calibration (during manufacturing) of energy meters can also cause NTL [18]. In most of the meter tampered locations, damaged meter terminals and/or illegal practices may not be visible during inspection.

![Diagram of a Meter System](image)

Figure 2-2: Technique used by illegal consumers to regulate the supply voltage and manipulate the energy meter reading [17].

### 2.2 Factors That Influence Illegal Consumers

Factors that influence consumers to steal electricity depend upon various local parameters that fall into multiple categories like social, political, economic, literacy, law, managerial, infrastructural, and economical. Of these factors, socio-economic factors influence people to a greater extent in stealing electricity. More concisely, some of the
important factors are:

- The belief that it is dishonest to steal something from a neighbor but not from a utility (public or large entity).
- Higher energy prices, unemployment or weak economic situation of a consumer.
- Corrupt politicians and employees of the utilities are responsible for billing irregularities. In some cases, total money spent on bribing utility employees is less than the money that would have been paid for consuming the same amount of electricity legally.
- Some consumers might not be literate about the issues, laws and offenses related to the energy theft.
- Weak accountability and enforcement of law.
- Reasons to hide total energy consumption (e.g. Consumers who grow marijuana illegally or small-scale industries to hide overall production/turnover).

In essence, electricity theft is proportional to the socio-economic conditions of the consumer.

### 2.3 Literature Review

In the recent past, several techniques were proposed for detecting the location of direct tapping on a feeder or tampered energy meter and identifying illegal consumers. On a parallel track, some non-technical measures, such as inspection of customers with
suspicious load profiles and campaigning against illegal consumption, were also implemented to control electricity theft. Some of the techniques (proposed worldwide) are described in this section.

A good strategy for fighting corruption in utilities has to be developed, considering the political scenarios, business processes, management techniques, and technologies in metering and distribution monitoring, control and automation based on the geographic location. In addition to the non-technical measures presented earlier, regularization of agricultural connections needs to be done. All of the contracts for deployment and maintenance of the distribution sector must be outsourced based on the performance of the enterprise to which the bid will be awarded. In addition, in most countries, electricity theft is considered a serious offense and illegal usage in any form of unbilled energy belonging to a utility is punishable under law. Laws and policies are being enforced such that political leaders do not protect corrupt employees and illegal consumers responsible for theft.

A constituency has been proposed to be created through effective communication with the important stakeholders, institutionalization of new business processes that adopt modern technology, and improvisation of management information systems [3]. Periodic inspection of illegal connections involves a lot of labor and strain for vigilant officials. The shunts detecting equipment proposed in [19] is time efficient and helps in the detection of electricity theft in underground distribution cables. Revenue Assurance and Audit Process (RAAP) is composed of macro-functions to detect and analyze revenues involved in illegal consumption of electricity. Also, Mano R. et al. suggests proper design and implementation of rules in the investigation of illegal consumers. RAAP is
targeted at improving the revenues for the utility by reducing commercial losses at about 20% each year [20]. In India, the Electricity Act of 2003 has made electricity theft a punishable offence and gave full freedom to vigilance officials to inspect and detect illegal consumers. In Pakistan, Karachi Electric Supply Company (KESC) has obtained a fatwa or decree, from Islamic scholars, declaring that illegal consumption of electricity is a sin [21]. On the other hand, teams are arranged for inspection and detection of illegal consumers of electricity, and their reward depends on the number of cases they inspect. Such incentives are proportional to the total number of illegal consumption cases they detect [22].

Several technical measures were also implemented in order to detect and help utilities in their battle against NTL. GE has patented an energy meter that reads electricity consumption correctly, even if the in-going and out-coming meter terminals are reversed. This invention stopped illegal consumers from using their energy meter in the reverse direction to reduce their utility bill. Installation of a prepaid energy meter can be a solution to monitor the distribution system and control electricity theft [23]. Location of electricity theft on a distribution feeder can be detected based on the values of the phase angle and impedance of the transmission lines at two different operating frequencies respectively [24]. Bandim C.J. et al. proposed utilization of a central observer meter at secondary terminals of distribution transformer. The value of energy read by the central observer meter is compared with the sum of energy consumption values read by all energy meters in range. These two values of the current are compared to estimate the total electricity that is being consumed illegally [25]. Vigilant Energy Metering System (VEMS) is a proposed energy metering system that can fight electricity theft. It has the
ability to collect, transfer and process data between other energy meters, local station and base station. It also identifies probable locations of theft and helps the utilities to control theft. A remote billing system can also be developed modifying this model [26]. Illegal consumption of electricity can be detected by using a remote check meter based on the amount of losses and time stamp of the check meter. This method is implemented before inspecting the illegal consumers personally by the vigilance officials, based on the data at the proper frequency of the consumer measurements [27]. A microcontroller based energy meter proposed by Jamil M. et al., gives utilities the ability to monitor and control the power supply of its spatially distributed consumers. This meter acts as a check meter that helps detect meters that have been tampered [28]. In addition, e-metering systems can collect and process data, as well as detect abnormalities in load profiles indicating electricity theft [29].

Nagi J. et al. proposed a novel approach of using Genetic Algorithm-Support Vector Machines (GA-SVM) for detecting illegal consumption of electricity. Load consumption data of all the households is collected, and data mining techniques are used to filter and group these customers before detecting illegal consumption. Customers are grouped into different classes based on the extent of the abnormality in load profile and customers with high probability of theft are personally inspected [30, 31]. The Extreme Learning Machine (ELM) approach is used to evaluate abnormal load behavior indicating electricity theft based on a load-profile evaluation. Nizar A.H. and Dong Z.Y. used online sequential-ELM (OS-ELM) algorithms in detecting and grouping the load profiles to reduce NTL [32, 33, and 34].
2.4 Summary

To summarize Chapter 2, methods implemented by illegal consumers in pilfering electricity are explained. Several setbacks in implementing existing technical measures are analyzed. Several factors that influence illegal consumers to steal electricity are explained. Relevant literature review of various methods and techniques proposed and implemented for controlling illegal consumption or identifying illegal consumers.
Chapter 3

Smart Meters for the Power Grid – Challenges, Issues, Advantages and Status

This chapter presents an extensive overview of the technologies involved in smart meter hardware and communication. Issues in the design, development and deployment of smart meters, and data management are also presented. This chapter discusses the features and advantages of smart grid and smart meters which enabled the proposed research (in this dissertation) in the real-world.

A smart meter is an advanced energy meter that measures the energy consumption of a consumer and securely communicates this and additional information to the utility. The ability of smart meters for bi-directional communication of data enables the utilities to collect information regarding the electricity fed back to the power grid from customer premises. In addition to secure communication, smart meters can execute control commands remotely as well as locally. Therefore, smart meters can be used to monitor and control all home appliances and devices at the customer’s premises. They are capable of collecting diagnostic information about the distribution grid, home appliances, and can communicate this information with other meters in their reach. They measure electricity
consumption, support decentralized generation sources and integrate energy storage devices. Distributed power generation sources would be an essential and integral part of future households. All of these additional services and demand management techniques require utilities to collect large quantities of real-time data.

Smart meters can be programmed such that only power consumed from the electricity grid is billed, excluding the power consumed from the distributed generation sources or storage devices owned by the customers. In addition, a customer will be credited for the energy supplied to the grid. Data communicated by a smart meter is typically a combination of parameters such as a unique meter identifier, timestamp, and instantaneous energy consumption values. Smart meters can limit the maximum electricity consumption, control home appliances, and can terminate or reconnect electricity supply to any customer remotely in case of a fault or event in the neighborhood [35], [36].

In a smart grid environment, smart meters play an important role in monitoring the performance of the grid and the customer energy usage characteristics. Collection of energy consumption data from all customers at regular intervals of time allows the utility to manage and optimize electricity demand more efficiently. Home energy management techniques with the help of smart meters also advise customers about the cost efficient operation of home appliances. In light of this, smart meters can be used to control light, heat, air conditioning and other appliances [37]. Smart meters can be programmed to operate home appliances in a defined schedule. In addition, integration of smart meters helps utilities in detecting unauthorized consumption and electricity theft in view of improving the distribution efficiency and power quality [38].
3.1 Communication Technologies

Utilization of the smart meter system involves a large quantity of data transfer between the utility, smart meter, and home appliances in the network. This data is sensitive, confidential, and access to this data should be given to only a few personnel. With the restrictions on this data, security guidelines are formulated for transmission, collection, storage, and maintenance of the energy consumption data. The communication standards and guidelines were formulated to ensure that data transfer within the network is secure. It is equally important that this data must represent the complete information regarding the customer’s energy consumption and status of the grids without any potential manipulations or miscalculations. So, this data must be authenticated and should reflect information about the target devices correctly [39]. Figure 3-1 shows the generic architecture of a communication network that is capable of performing all the features discussed above. In this figure, devices in the transmission sector ensure proper transmission of generated energy, control systems in the distribution sector ensure monitoring and controlling of faults, communication devices like protocol gateways, data collectors, repeaters and network operations coordinate data as well as control signals between all the devices in the communication network.

The common network selected has to support the required operation of the smart meter system even on power outage and support distribution automation. In addition, the selected network and its components must be cost effective and must support “traffic prioritization” i.e. they must prioritize the delivery of data based on its time and direction sequence [40]. Communication technology to be chosen should be cost effective, provide
good transmission range, better security features, bandwidth, and power quality with least possibility of repetitions.

Bluetooth technology can be a possible option for communication of control signals and transmission of energy consumption data. In view of implementing this technique, B.S. Koay et al. proposed a Bluetooth based energy meter that can collect and transmit the energy consumption data wirelessly to a central base station [42]. Power Line Communication (PLC) and Broadband over Power Line (BPL) communication are
the other possible options of data transfer supporting the higher level communication suites such as TCP/IP. PLC uses the existing electricity grid, cellular/pager network, mesh network, a combination of licensed and unlicensed radios, wireless modems, existing internet connections, power line communications [43], RS-232/485, Wi-Fi, WiMAX, and Ethernet with a protocol to upload data using IEC DNP [44]. PLC technology effectively automates the process of data collection in smart meter applications [45]. Despite huge overhead due to IPv6, IPv6 can be applied to physical layer with lower data rates. However, IPv6 combined with Media Access Control (MAC) algorithm accomplishes less delay time and higher throughput. Though this combination might slightly reduce the usable data transfer rate, it will not affect the overhead at the MAC layer [46]. IP based network protocol could be another promising option for communication because of its advantages over other technologies while satisfying the security standards of the smart grid communications. In addition, TCP/IP forms an efficient communication platform across multiple devices [47].

In addition, Session Initiation Protocol (SIP), a text-based signaling protocol, is employed for controlling multimedia sessions like video and Voice Over Internet Protocol (VoIP). SIP integrates several features of HTTP and Simple Mail Transfer Protocol (SMTP). SIP is an open and standards-based technology, which provides a robust communication medium for the smart grid applications [48]. SIP can be implemented on top of TCP, User Datagram Protocol (UDP), or Stream Control Transmission Protocol (SCTP).

A new architecture based on DNP3 is proposed by T. Mander et al. DNP3 produces a protocol discontinuity between DNP3 devices (used for regulated power
system operations) and TCP/IP devices (used for the smart load and demand management). The advantage with this architecture is, the discontinuity limits the vulnerable attacks from other TCP/IP devices. Some security enhancements such as data object security and a security layer may be added to DNP3, as this protocol by itself is not adequately safe for collaborative operations. Data object security appends additional rules to access data thereby preventing the unauthorized access that can potentially manipulate data and device operations [49].

An energy meter based on Peer-to-Peer (P2P) network is presented in S. Rusitschka et al. The utilization of P2P network enhances the range of operations. In addition, several value added services can be employed. P2P communication uses the internet, which leads to a cost effective design of smart grid communication networks [50]. In addition, the P2P network utilizes the resources of participating homes optimally.

Yet another network, Zigbee [51], is a potential communication network for transfer of data as well as control signals. As many industrial and household entities maintain a computer with 802.11.x, Zigbee protocol can be used with Home Area Networks (HANs) for data transfer over 802.11.x [52]. This technology can be used instead of increasing the operating clock frequency in the crypto core in order to reduce the response time and verification delay; J. Kim et al. proposed the mode toggling approach on the design process for AES-CCM module. They have also adopted the optimal security material management module. These design methodologies and the obtained response time allow the cryptographic core to maintain the minimum clock frequency, while staying within the constraints, ensuring the reduction in total dynamic power consumption [53].

20
General Packet Radio Service (GPRS) technology is another potential communication medium for transferring both the data and control signals wirelessly over long distances. In contrast to other communication network technologies, only a few communication characteristics that represent GPRS communication network have been assessed. That being said, lack of tools for detecting a network failure would be a major setback in implementing GPRS network in many geographical locations. Before deploying a GPRS based communication system in a specific location, availability and quality of the signal has to be determined [54]. Parallel processing and implementation of the Field-Programmable Gate Array (FPGA) hardware can reduce the time elapsed for interpreting the data and obtaining the status of the distribution network. Adoption of reconfigurable logic for processing of data minimizes the amount of data to be generated by a smart meter [55].

3.2 Issues and Challenges

In general, efficient management of the grid can be an alternative solution instead of revamping the existing grid. Considering the technical advantages and enhancements in the operation capability, integration of the smart grid stands as a valuable solution in managing the existing grid. However, the design, deployment, and maintenance of the smart meter system involve many issues and challenges. Deployment and maintenance of smart meter system in a distribution grid involves several billion dollars of investment. So, this investment has to be realized for the projected increase in the energy demand and distributed generation [56]. Initially, the process of replacing the existing energy meters
with a smart meter system will be a challenge for utilities. Lack of a proper infrastructure for synchronizing this new technology with the existing ones might interrupt the introduction of smart meters. Though several devices are integrated with the smart meter system, they can be used to their fullest extent only when all the appliances and devices in the distribution and metering network are included in the communication network. Integration of these devices becomes further complicated with an increase in the number of customers. Deployment of communication networks in some localities might also be difficult due to terrestrial difficulties [47]. In the USA, utilities receive incentives for selling more electricity, which might not drive them to encourage their customers to conserve energy [35].

Collection and transmission of energy consumption data is a continuous process that needs to be done automatically, which is a tedious and expensive job. In this context, a common notion might arise among several customers that smart meters might create some privacy and security risks as the data and signals are being transmitted. Additionally, this data might also reveal information about the presence of people at their residence, when they were present, and what appliances were being used. In view of this, some customers might be unwilling to communicate their energy consumption data with their neighbor’s meter. Fundamentally, it would be an issue of determining which parameters should be transmitted and who is authorized to access this information [52], [57-60].
3.3 Summary

To summarize Chapter 3, a detailed analysis on the features of smart meters and available communication technologies are presented. Issues and challenges in the design, development, deployment, and maintenance of smart meter technologies are explained.
Chapter 4

Analysis and Development of Energy Consumption Patterns

4.1 Introduction

This chapter explains the importance and need for analyzing energy consumption data and features of the energy consumption data for the identification of illegal consumers. The most feasible way of identifying illegal consumers is by analyzing the data or power system parameters that represent a situation at the customer end. Instantaneous customer energy consumption is the most significant parameter that best represents irregularities in customer energy consumption. In addition, analyzing customer energy consumption patterns is one of the essential mechanisms carried out by the utilities for clear understanding about the condition of grid at the customers end.

4.2 Generating Energy Consumption Data

With the advent of smart meters and other smart grid infrastructure, it is now
possible to access, collect, and analyze instantaneous energy consumption of customers in real-time as explained in chapter 3. Exploiting such features offered by smart grid, customer energy consumption data is used to classify or identify illegal consumers. However, the energy consumption data required to test the proposed algorithms is unavailable owing to the privacy and confidentiality of utilities and customers. Therefore, this data has been developed using MATLAB in the following procedure.

The required data has been carefully developed to match very closely with real-world data. Data about the hourly electricity load on a distribution feeder has been obtained from PJM datasets [61]. This data represent energy consumption of a feeder supplying a large geographic entity or a few neighborhoods. Therefore, it has been considered that customers on the feeder include several residential, small industrial and commercial customers. There may be several customers in each of the mentioned categories in every neighborhood. These values of overall load on the grid have been divided among these customers, after assigning a range of energy consumption for each type and range of customers. To this end, all categories of customers on the grid are grouped based on their energy consumption over a period of time. In this analysis, parameters considered for the division are:

- Season of year (summer, winter, and the rest of the year)
- Types of customers
  - Agricultural (small, large),
  - Commercial (small, medium, large),
  - Residential (small, medium, large)
Table 4.1 Types of agricultural customers based on the ranges of energy consumption.

<table>
<thead>
<tr>
<th>Consumption</th>
<th>Small</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>kWh per month</td>
<td>0-100</td>
<td>100-300</td>
</tr>
<tr>
<td>kWh per day</td>
<td>0-3.333</td>
<td>3.333-10.0</td>
</tr>
<tr>
<td>kWh per hour</td>
<td>0-0.14</td>
<td>0.14-0.42</td>
</tr>
</tbody>
</table>

Table 4.2 Types of commercial customers based on the ranges of energy consumption.

<table>
<thead>
<tr>
<th>Consumption</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>kWh per month</td>
<td>0-500</td>
<td>500-2000</td>
<td>2000-20000</td>
</tr>
<tr>
<td>kWh per day</td>
<td>0-16.677</td>
<td>16.677-66.677</td>
<td>66.677-666.677</td>
</tr>
<tr>
<td>kWh per hour</td>
<td>0-0.695</td>
<td>0.695-2.787</td>
<td>2.787-27.877</td>
</tr>
</tbody>
</table>

Table 4.3 Types of residential customers based on the ranges of energy consumption.

<table>
<thead>
<tr>
<th>Consumption</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>kWh per month</td>
<td>&lt;300</td>
<td>300-600</td>
<td>&gt;600</td>
</tr>
<tr>
<td>kWh per day</td>
<td>1.667-10</td>
<td>10-20</td>
<td>20-33.333</td>
</tr>
<tr>
<td>kWh per hour</td>
<td>0.069-0.416</td>
<td>0.416-0.833</td>
<td>0.833-1.38</td>
</tr>
</tbody>
</table>

It has been considered that each neighborhood may consist of multiple customers representing some or all of the categories (of customers) illustrated above. Therefore, load on each neighborhood has been divided among these customers in several possible combinations (number of customers per each type and ranges of types). As a result, about 20,000 average consumption patterns representing energy consumption of 20,000 customers over a day has been developed. These energy consumption patterns feature instantaneous energy consumption collected 96 times a day. In other words, energy consumption readings from smart meters are assumed to be collected in real-time from 20,000 customers at 15 minute intervals.

Irregularities representing electricity consumption by illegal consumers must also be incorporated into some data samples of the developed data. Therefore, energy
consumption readings of some customers at some timestamps have been reduced, and some readings are replaced by zero. These modifications are carried out on energy consumption data samples belonging to random customers and at random instances of time. Zero represents no-consumption and modified energy consumption represents partial legal consumption. As a result, the process of developing energy consumption data is complete. In the final dataset, some of the energy patterns demonstrate the consumption by genuine customers, some partial illegal consumers and the rest represent complete illegal consumers.

4.3 Validation of Energy Consumption Data

As the consumption of electricity by a customer depends on their requirements, energy consumption patterns of various customers are different. In addition, weather is one of the most important factors that influence the electricity consumption. Therefore, energy consumption data may be considered as a function of size of the customer, geographic location, season of the year, weather conditions, and time of day.

As an example, two of the developed energy consumption patterns are illustrated here. Figure 4-1 shows the approximate energy consumption of a small residential customer on a weekend. In this example, most of household electrical appliances may be used during the day-time in summer. Day-time electricity consumption is slightly higher than at night, as only the basic electrical appliances may be used. In the same geographical location, for a large customer, the energy consumption may be high during the day (compared to night) due to the operation of large air-conditioning units. Figure 4-
2 demonstrates the approximate energy consumption pattern of a different small residential customer on a weekday. During summer, energy consumption is much higher than the rest of the year as air-conditioning units and refrigerators may be operated more.

Figure 4-1: Instantaneous energy consumption pattern of a residential customer on a weekend for three seasons.

Figure 4-2: Instantaneous energy consumption pattern of a different residential customer on a weekday for three seasons.

Small customers employ equipment with light loads to meet their basic needs, and these customers report very low consumption during night. On the other hand, large commercial customers employ heavy loads including large air-conditioning units and large storage units that consume large quantities of energy. So, typically the energy
consumption will be very high for maintaining the room temperature on the premises and therefore consumption may be less in winter.

Based on the developed energy consumption patterns, consider a neighborhood with multiple small, medium and large residential, and small, medium and large commercial customers. Figure 4-3 illustrates the energy consumption of that neighborhood with distribution losses of 30.9% on a summer weekday. In Figure 4-3, real-time energy consumption readings portray the energy read by all energy meters in the neighborhood. From the known value of total energy supplied, the energy loss curve can be mapped. Therefore, the average value of total energy supplied over a 24-hour period is 489.029 kWh and the energy lost is 151.121 kWh.

![Figure 4-3: Overall load on feeder for a neighborhood with total losses of 30.9 % on a summer day.](image)

To simplify the analysis, let us consider energy consumption of one small residential customer on a weekday during a 24-hour day period as shown in Figure 4-4. Figure 4-5 presents hypothetical scenarios of energy consumption readings of the same customer for the following three cases:
1. Genuine consumption: if a customer consumes entire energy for the household legally.

2. Partial illegal consumption: if a customer consumes a portion of required household energy legally and the rest illegally.

3. Complete illegal consumption: if a consumer steals the entire portion or a majority of the total household energy illegally.

Figure 4-4: Approximate energy consumption pattern of a small customer on a weekday in summer.

Figure 4-5: Energy meter readings collected from a consumer’s smart meter, projected for three modes of consumption - genuine consumption, partial and complete pilfering.
For implementation and validation of the proposed encoding and detection (classification) algorithms, the energy consumption patterns (developed) for 20,000 customers have been used. The overall energy consumption patterns of these customers are similar to Figures 4-1, 4-2, and 4-4 in terms of the total number of energy readings (96 inputs), but varies in the amount/magnitude of energy consumption. The developed data has been used to ensure whether or not the proposed detection algorithm detects the difference and uniqueness between energy consumption patterns of customers, analyze their consumption, and group them into different classes based on irregularities. In essence, the proposed algorithm captures the pattern of energy consumption with reference to the overall load on the grid.

4.4 Novel Encoding Scheme for Energy Consumption Data

In the real-time operation of algorithms for detection of illegal consumers, instantaneous energy consumption data can be analyzed. Algorithms used for the detection of illegal consumers tries to capture pattern of energy consumption of a customer, and classify that customer based on irregularities in consumption. Furthermore, instead of inputting energy consumption data, irregularities or discrepancies in customer energy consumption can be directly inputted to the detection algorithms for enhanced identification. An encoding technique has been developed to map or modify energy consumption patterns into irregularities. However, this encoding process should result in the following advantages: faster and easier analysis, quicker classification, lesser space
for storing historic data, better understanding of customer energy consumption, preserve quality and uniqueness of the data. Figure 4-6 illustrate the first stage of the encoding.

Figure 4-6: First stage of the encoding process – this stage simplifies the energy consumption data.
To simplify and modify energy consumption data of customers, the following encoding procedure has been adopted: First, inputs or data points corresponding to a customer (energy consumption reading of that customer’s energy meter) with a zero value (zero energy consumption) are considered as ‘0’ and inputs with non-zero energy consumption are replaced with ‘1’. Energy consumption reading of all customers and at all points of time should be modified. After making sure that all values are either ‘1’s or ‘0’s, three consecutive inputs are considered as a group. After grouping, the initial 96 inputs are reduced to 32 sets of energy consumption data corresponding to each customer. Then, in each of these sets, the first input (most recently received) is multiplied by ‘4’, the second input by ‘2’, the third input by ‘1’, and these three multiplied values are added. This addition should be completed for all consecutive groups (of three inputs) and for all customers. At this stage, the number of inputs corresponding to each customer is 32. As a result, all inputs in the encoded dataset have become whole numbers in the range of ‘0’ and ‘7’. This concludes the first stage of the encoding process.

In the second stage of encoding, six different time zones have been identified. These time periods are divided such that energy consumption of a typical customer will not significantly vary within those intervals. Then, a random reading in each of these time–periods is selected, as displayed in Table 4.1. During the first week of every month, overall energy consumption of the neighborhood as well as the instantaneous energy consumption patterns of every consumer at those particular instances of time have to be collected (recorded) for three days in a row. These three energy consumption readings and overall load values (collected over three days) will be averaged and stored. Individual energy consumption values and total load on the feeder of that neighborhood have been
collected for inputs, ‘16, 28, 36, 56, 72, and 88’ in regular data (corresponds to input ‘5, 9, 12, 18, 24, and 29’ in encoded data). These values will serve as a reference for energy consumption of a particular customer in their respective time zones.

Table 4.4 Time zones and corresponding inputs.

<table>
<thead>
<tr>
<th>Number</th>
<th>Time Zone</th>
<th>Considered input</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12:00 AM – 06:00 AM</td>
<td>16 (04:00 AM)</td>
</tr>
<tr>
<td>2</td>
<td>06:00 AM – 08:00 AM</td>
<td>28 (07:00 AM)</td>
</tr>
<tr>
<td>3</td>
<td>08:00 AM – 10:00 AM</td>
<td>36 (09:00 AM)</td>
</tr>
<tr>
<td>4</td>
<td>10:00 AM – 05:00 PM</td>
<td>56 (02:00 PM)</td>
</tr>
<tr>
<td>5</td>
<td>05:00 PM – 08:00 PM</td>
<td>72 (06:00 PM)</td>
</tr>
<tr>
<td>6</td>
<td>08:00 PM – 12:00 AM</td>
<td>88 (10:00 PM)</td>
</tr>
</tbody>
</table>

After the average energy consumption values are calculated, a variable called the Multiplication Factor (MF) has been introduced. Here, MFs help estimate the energy consumption of a specific customer at a specific point of time. MFs are calculated for all customers at those instances of time (inputs) mentioned in the Table 4.1. MF is the ratio of average energy consumption of a consumer at a particular time (inputs, as in Table 4.1) to the total energy consumption of neighborhood. MFs calculated for a particular instance of time in each time-zone represents MF for that entire time-zone. MF multiplied by total power on the grid (for the neighborhood where MFs are calculated) gives an estimate of energy consumption by a specific consumer at a given point of time. Then, the probable energy consumption of each customer has to be calculated in the following procedure: Original, instantaneous consumption of all customers (with 96 inputs) is considered; this data is then grouped into sets of three consecutive instantaneous energy consumption readings (inputs at three consecutive timestamps); therefore, the customer’s
real-time energy consumption is grouped into 32 sets. Averages of the three energy consumption readings in each of the sets are calculated. These averaged values are termed as Average Original Consumption (AOC). Then AOC’s of all customers (32 AOC values per customer) are considered and added individually to obtain a single data sample, termed as Average Total Consumption (ATC) with 32 inputs for a single day. Therefore, ATC represents energy consumption of a set of customers in a neighborhood or at feeder level. Now, probable energy consumption i.e. Calculated Probable Consumption (CPC) of a particular customer at a particular point of time has been calculated using MFs as shown in (4.1).

\[
CPC = \frac{(ATC*MF)}{Number\ of\ Customers}
\]  

(4.1)

The real-time energy consumption data (for testing) modified as per the first stage of encoding has to be further modified considering CPC. This step ensures inclusion of the effect of partial illegal consumption. This additional modification of the energy consumption data can be carried out in the following procedure: At any point of time, instantaneous energy consumption of every customer is assumed to be 80% or more of the CPC (calculated for that customer). However, if it is less than 80% of CPC at any point of time, then the encoded energy consumption data should be modified as follows: At that point of time during which the original consumption is low (most recent or latest input in the set of three inputs), if a particular consumer’s instantaneous energy consumption is less than 90% of CPC, then the corresponding encoded energy consumption (input) will be deducted by ‘2’; if a particular consumer’s instantaneous energy consumption in an earlier time-step (second input in the set of three inputs) is less than 90% of CPC, then the corresponding encoded input will be deducted by ‘1’; if a
particular consumer’s instantaneous energy consumption in the second earlier time step (first input in the set of three inputs) is less than 90% of the CPC, then the corresponding encoded input will be deducted by ‘1’. In essence, due to the second stage of encoding, value/magnitude of several inputs (encoded energy consumption values) of different customers may be slightly less when compared to values of the previous stage of encoding. Figure 4-7 illustrates the second stage of the encoding process. In essence, MF is calculated before starting the testing procedure and AOC, ATC, and CPC are calculated while classifying the customer energy consumption data. Note: percentages (80% and 90%) in the encoding algorithm are chosen optimally, such that a genuine customer is not classified as a customer responsible for partial illegal consumption. Figure 4-8 displays instantaneous energy consumption of a random customer over a single day. Zero consumption around the 10th reading in the figure may indicate illegal consumption. Large variations in energy consumption with respect to the time of the day may be seen as demand varies over time.

Figure 4-9 illustrates the instantaneous energy consumption of all 440 instances (customers) at a specific point of time. Figure 4-10 displays calculated AOC of all customers at a specific time. Figure 4-11 displays the MF calculated for the first 100 customers at a specific point of time. Figure 4-12 displays another set of MF values for the same 100 customers, at a different point of time. Figure 4-13 illustrates encoded energy consumption of 200 customers at a particular instance of time (one input of 32). It can be observed that the magnitude of encoded energy consumption ranges between 0 and 7. Figure 4-14 displays a detailed form of the inputs (energy consumption) for first 20 customers after second stage of encoding.
Figure 4-7: Modifications done as a part of second stage of the encoding process – this stage modifies the result from first stage of encoding.
Figure 4-8: Instantaneous energy consumption of one customer over an entire day – an example extracted from the developed data.

Figure 4-9: Instantaneous energy consumption readings of 440 customers recorded at a specific time. Large variations represent wide range of customers on grid.
Figure 4-10: Calculated average energy consumption values for 440 customers at a specific time, as a part of the encoding process.

Figure 4-11: MFs calculated during the encoding process for 100 customers at a specific time.
Figure 4-12: MFs calculated during the encoding process for a different set of 100 customers at a specific time.

Figure 4-13: Irregularities of 200 customers at a specific time obtained as a result of the encoding process.
Figure 4-14: Irregularities of 20 customers at a specific time obtained as a result of the encoding process.

Table 4.5 displays instantaneous energy consumption readings a customer received over a single day. A total of 96 energy readings can be seen with respect to time and corresponding serial number of the reading. Table 4.6 displays the equivalent irregularities. As can be seen the number of irregularities is 32 ranging between 0 and 7.

Table 4.5 Instantaneous energy consumption readings received for one customer in a single day.

<table>
<thead>
<tr>
<th>Energy</th>
<th>0.5</th>
<th>…..</th>
<th>0.6</th>
<th>0.6</th>
<th>……..</th>
<th>0</th>
<th>0</th>
<th>0.3</th>
<th>……..</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>00:45</td>
<td>…..</td>
<td>05:30</td>
<td>05:45</td>
<td>……..</td>
<td>13:15</td>
<td>13:30</td>
<td>13:45</td>
<td>…</td>
<td>24:00</td>
</tr>
<tr>
<td>Number</td>
<td>3</td>
<td>22</td>
<td>23</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.6 Encoded readings or pattern of irregularities generated by the encoding process.

<table>
<thead>
<tr>
<th>Irregularities</th>
<th>7</th>
<th>…..</th>
<th>4</th>
<th>…..</th>
<th>0</th>
<th>6</th>
<th>…..</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>00:15-00:45</td>
<td>…..</td>
<td>05:30-06:15</td>
<td>…..</td>
<td>12:45-13:30</td>
<td>13:45-14:30</td>
<td>…..</td>
<td>23:00-23:45</td>
</tr>
<tr>
<td>Number</td>
<td>1</td>
<td>8</td>
<td>18</td>
<td>19</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4-15 presents a comparison between the instantaneous energy consumption readings received from smart meters and CPC calculated for that customer for a single day. The magnitude of difference between these two values indicates an irregularity and the impact of this magnitude can be clearly seen in Figure 4-16. Figure 4-16 presents the pattern of irregularities for one customer in one day. Using the above mentioned encoding process, pattern of instantaneous energy consumption is mapped into pattern of irregularities as shown in Figure 4-16.

![Figure 4-15: Comparison between a customer’s instantaneous energy readings received and CPC calculated for a single day.](image-url)
In essence, first stage of encoding reduces the number of inputs to one-third and maps the energy consumption readings into irregularities in consumption, whereas in the second step, number of inputs remain unchanged, in the data will not be different, but the magnitude of inputs will be further be changed to represent partial illegal consumption. Fig. 3 illustrates the entire encoding process. This concludes the encoding process. After completing the encoding process for the existing set of instantaneous energy consumption readings (96 inputs), the encoding process has to be continued after receiving the next set of energy readings from smart meters in the next time-stamp. When the next data point (energy reading) of a particular customer is received, the number of inputs will be 97. Therefore, the earliest encoded input or the earliest set of 3 inputs (of 96) should be neglected or removed from the set of encoded data to be evaluated for detection of illegal consumers. Then, this newly received input should be duplicated twice such that it forms
a new set of three values (for encoding). Here, all three values in that set are the same. Therefore, the newly received reading will be considered as a set of three readings with the same value. After this modification, the entire encoding process should be carried out normally. In the next time-stamp, i.e. after receiving the next energy consumption reading from smart meters (second input in the set of three), then the average of first two readings in the set will be considered as third value in the set. The encoding process will be performed with this new set of energy consumption readings. Finally, when the next real-time energy consumption readings are received from smart meters, set of three readings are available. Now, the encoding process can be performed with this new set of readings.

4.5 Summary

To summarize Chapter 4, importance of detecting and grouping illegal consumers by analyzing customer energy consumption patterns is discussed. Procedure adopted in developing near real-time approximate energy consumption data for a wide range of customers is illustrated. A novel data encoding technique is proposed to simplify the analysis of energy consumption patterns and map them into patterns of irregularities in energy consumption by each customer. Detailed procedure involved in implementing the encoding technique is presented. In addition to the faster analysis, the encoding technique enhances the entire process of detecting illegal consumers.
Chapter 5

Identifying Illegal Consumers using Intelligent Classification

This chapter illustrates the algorithms implemented for detecting illegal consumers. These classification algorithms include SVM, Rule Engine, and Neural Network Pattern Recognition (NNPR) tool based classification models. The classification results of the proposed classification algorithms are presented.

5.1 SVM Based Classification Model

SVMs introduced by Vapnik are a set of supervised learning methods. They can analyze the given data and recognize a pattern or trend in the data with respect to output. SVMs are also used for regression analysis and statistical data classification. Given a training dataset that represent a set of rules, a model can be developed by the SVM using a training algorithm [62], [63]. In general, SVMs develop a hyper plane or set of hyper planes in a high or infinite dimensional space, depending on the complexity of the data that needs to be classified. Significant separation between the classified data points can be achieved when the hyper plane has significant distance to the nearest training data
points of any class. The generalization error of the classifier will be minimal if the separation margin is high. In the recent past, SVMs have found numerous applications in face recognition, text categorization to bioinformatics, and data mining [64]. The training data with \( x_i \in \mathbb{R}^n, i = 1, \ldots, l \), in two classes, and a vector \( y \in \mathbb{R}^l \) such that \( y_i = \{1, -1\} \), C-SVC [62], [65] is used to solve

\[
\min_{w, b, \xi} \left[ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \right]
\]

Subject to

\[
y_i (w^T \varphi(x_i) + b) \geq 1 - \xi_i, \quad (5.2)
\]

\[
\xi_i \geq 0, i = 1, \ldots, l. \quad (5.3)
\]

Here, LibSVM [66] is used for developing the required classification model. LibSVM is a library for developing SVMs based on classification model in MATLAB developed by C.C. Chang and C.J. Lin. In power engineering, SVMs are used for several applications including estimation of electricity theft and analysis of power quality parameters in a power grid. Classification accuracy is the ratio of correctly classified data samples over all data samples. The customers are classified into three classes based on the following criteria and their instantaneous energy meter readings (rules):

To be classified into Class-D: in instantaneous energy meter readings of any customer,

- If zero energy consumption is recorded for more than two hours in one day or
- If zero-energy readings are recorded more than 8 times (eight of 96 consecutive readings) including repetitions.

To be classified into Class-S: in instantaneous energy meter readings of any customer,

- If 3 readings (of 4) in any hour of a single day are recorded as zeros
If two consecutive zero readings are recorded in an entire day with less than three repetitions,

If zero energy readings are recorded between three to six times in a day.

To be classified into Class-I: in instantaneous energy meter readings of any customer,

If zero energy readings are recorded less than two times in one day including repetitions.

Figure 5-1 illustrates the operational flowchart of the SVM classification algorithm. Initially, utilities collect instantaneous electricity consumption data from the smart meters in specific intervals of time. Thus, collected energy consumption data would be a series of instantaneous energy consumption values. Format of the electricity consumption data collected from smart meters need to be modified; so that, it would be compatible with SVM model developed using LibSVM. A portion of the data developed in chapter-4 has been extracted as training data and the rest as testing data. Inputs to the SVM model is the instantaneous energy consumption data, and the output is the classes that a particular customer belongs to. Before the data being used for training and testing, it may be viewed categorically based on the geographical location, whether it is a weekday or weekend, load capacity range of the customer and what season of the year this data represent. The data is then transferred to a database located at a central control station. Then, training data is used to train the SVM model and test it for detecting the illegal consumers. If a customer profile is genuine and the energy consumption is continuous, then that the customer is rated as a genuine customer. If the customer profile is suspicious, then the profile needs to be evaluated further.
Figure 5-1: Operational flowchart of the proposed SVM based classification model.

If a customer’s energy consumption fulfills criteria specified for Class-D, then, the customer may be inspected immediately, as the probability of illegal consumption is very high. If a customer is classified as Class-S, and if the customer is either a large or medium customer then that customer is immediately inspected. If that customer is a small
customer, then that customer is periodically inspected. If a customer falls under Class-I, that customer can be reported as a genuine customer. If a customer’s profile does not fall under any class and the calculated overall distribution losses are more than 4% (excluding the classified illegal consumers), then the customer profile is reevaluated. In general, distribution losses are ideally to be between 3–5% at feeder level. If the losses are more, then it can be assumed that illegal consumers might exist on the distribution feeder. Therefore, classification algorithm may be terminated if the losses are less than 4%, but if the losses are more than 4%, the classification algorithm is reiterated.

The classification model developed using LibSVM needs a set of following parameters: probability estimate, a value of $\gamma$, cost parameter and training function. Values of $\gamma$ and cost function are selected to obtain minimum generalization error. In this work, data is trained with a probability estimate of ‘1’ and the pre-computed kernel function. Evaluation has been performed for $\gamma$ values ranging from 0.07 to 25. The best classification is obtained for a $\gamma$ value of 24.24, probability estimate of ‘1’, and radial bases function as the kernel. The accuracy of the classified instantaneous energy consumption data is 98.07% when a testing data for 880 customers/instances with 96 inputs. Overall, time consumed for training and testing of this model consumed 2.34 seconds. When, the SVM model is applied on a larger data set, i.e. Training data: number of customer energy consumption patterns: 440; Testing data: number of customer energy consumption patterns: 16200. Average classification accuracy for 20 different combinations of model parameters is 92.8%. Average testing time for each classification is 1.9 seconds. The values of $\gamma$ range between 0 and 0.6. CPU time for classification is less compared to the previous attempt for classification because of a lower range of $\gamma$.
values. When applied on a dataset with Training data – number of customer energy consumption patterns – 12000; testing data – number of customer energy consumption patterns – 16200. Classification parameters: probability estimate – 1, Radial basis function, \( \gamma = 0.1 \), Classification accuracy – 90.5\% (14661 correctly sampled of 16200), training and classification time – 158.74 seconds.

Figure 5-2 shows the classification results - the output classes mapped with respect to the customer identification number. Customers classified as ‘1’ comes under class-I, customers classified as ‘0’ are rated as class-S, and customers classified as ‘-1’ are rated as Class-D. Customers classified under Class-D are supposed to have stolen electricity and/or responsible for billing irregularities. Customers classified under Class-S are monitored for more time and checked for any repetitions in suspicious behavior. Based on the other rules customers can be placed under Class-D. Figure 5-3 is plotted for original output classes and predicted classes of output against the number of customers. Figure 5-4 illustrates the output classes predicted with respect to two random input values from all instances of the testing data. In this figure, the output classes ‘0’s are marked as ‘o’, ‘1’s are marked as ‘+’, and ‘-1’s are marked as ‘*’. Energy consumption data illustrates that most of data points fall into two clusters. The major problem in analyzing the energy consumption data is that, customers of same range (size) in a particular geographic area may consume relatively same magnitude of energy. Figure 5-5 shows a portion of the lower cluster in the range of \([0, 1]\) and \([0, 1]\). Figure 5-6 shows the probability estimates of three classes of the testing data.
Figure 5-2: Graph that illustrates the output classes predicted by the SVM classification model with respect to customer ID.

Figure 5-3: Classes predicted by the SVM classification model plotted against actual classes (of consumers) and customer IDs.
Figure 5-4: Energy consumption readings of one customer mapped against another random customer.

Figure 5-5: Detailed view of the lower cluster of data points in the Figure 5-4.
Figure 5-6: Probability that customers are classified into three output classes – represented using different colored lines.

Classification results of the customer energy consumption data after the encoding process are presented. Classification has been performed based on the following rules.

To be classified into Class-D: in encoded energy meter readings (inputs) of any customer,

- Two or more inputs should be zero and/or
- Six or more inputs are any integer between 0 and 6.

To be classified into Class-S: in encoded energy meter readings (inputs) of any customer,

- Not more than one input can be ‘0’ and/or
- Not more than four inputs integer between 0 and 6.

To be classified into Class-I: in encoded energy meter readings (inputs) of any customer,

- Not more than three inputs are numbers between 1 and 6.

After implementation of the encoding procedure, the obtained results both from regular
and encoded data are illustrated after consumption:

**Case 1**: Classification of 440 customers with 96 inputs each and for $\gamma = 0.34$:

- Classification accuracy of normal data: 93.5%.
- Classification time for normal data: 0.09 seconds.
- Classification accuracy of encoded data: 99%.
- Classification time for encoded data: 0.03 seconds.

**Case 2**: For classifying the 440 customers’ energy consumption data for 1692 different combinations of SVM model parameters and testing functions:

- Best classification accuracy of normal data: 93.5%.
- Classification time for normal data: 1508.654 seconds.
- Best classification accuracy of encoded data: 99%.
- Classification time for encoded data: 381.6638 seconds.

In pattern recognition (for identifying the illegal consumers), identification of energy consumption pattern is the priority than particular values in energy consumption patterns. In essence, encoding technique has the following advantages: the original consumption pattern is preserved, the problem size reduces, but the scope of the problem is not compromised. In addition, the time required for classification has become much smaller. In essence, time taken for classification of encoded data is about 25% of the time required for classifying the instantaneous energy consumption data, and superior classification accuracy can be achieved for small values of $\gamma$.

Figure 5-7 illustrates the CPU times recorded for classifying normal data and encoded data against several attempts of classification. Here, the total time taken for classification of all inputs is 1508.65 seconds and the average time per each classification
is 0.89 seconds. Figure 5-8 illustrates the classification time taken for encoded data, and Figure 5-9 illustrates the classification time taken for normal data with respect to number of attempts. Here, the total time taken for classification of all encoded inputs is 381.6638 seconds and the average time per each classification is 0.22557 seconds. In another attempt of classification, the total time taken for classification of all inputs for normal data is 104.11 seconds and the average time for each classification is 0.043 seconds. The total time consumed for classification of encoded data is 29.46 seconds, and the average time for each classification is 0.38 seconds. The SVM model classification results after the encoding are presented here.

![Classification Time Graph](image)

**Figure 5-7:** CPU times for different attempts of classification by the SVM classification model for normal and encoded energy consumption data.

Figure 5-10 illustrates classification accuracy of the SVM model for different combinations of model parameters. Classification accuracy of the SVM model ranged between 76–92%. Figure 5-11 displays classification times (SVM model CPU time) for multiple iterations. The classification time for obtaining the best accuracy (92%) was 0.1
second (including the time consumed for encoding). These results demonstrate the performance of SVM model for classification of encoded energy consumption readings. Figure 5-12 maps classes predicted by SVM model vs. original classes of testing data.

![Graph](image)

**Figure 5-8:** CPU times for different attempts of classification by the SVM classification model for encoded energy consumption data.

![Graph](image)

**Figure 5-9:** CPU times for different attempts of classification by the SVM classification model for normal energy consumption data.
Figure 5-10: Classification accuracies of the SVM based classification model for different attempts of classification (different sets of model parameters).

Figure 5-11: CPU times for different attempts of classification by the SVM classification model for different attempts of classification (sets of model parameters).
5.1.1. Neural Network (NN) Model

Training and testing the SVM model using LibSVM requires us to select the function to be used for training the model, parameters involved in that training function, probability estimates, and cost function as mentioned earlier. These parameters and training function used to train the SVM model depends on the type, size, and pattern of the data. The real-time energy consumption data varies depending on the size of the customer, geographical location, time of the year, and type of the customer – the pattern of energy consumption data varies dynamically over time. Therefore, the function and set of parameters used to train the SVM model may need to be changed accordingly. As a result, it is a difficult and time-consuming process to train and test the SVM model with multiple sets of parameters before choosing the ideal set. Figure 5-13 presents the variation of SVM model classification accuracy plotted against $\gamma$ value for probability
estimate of ‘1’. Whereas, Figure 5-14 presents the variation of SVM model classification accuracy plotted against γ value for probability estimate of ‘0’. From the pattern in which the classification accuracy of the data varies with γ, it is observed that, CPU time consumed for classification increase with an increase in γ value. To this end, a NN model has been developed to estimate the required set of parameters and respective functions to achieve good classification accuracy and save time in real-time operation. A NN is an efficient tool to detect and classify the patterns appropriately. In a NN model, the knowledge is stored as the strength of interconnections (weights) [67] between neurons. The classification problem guides the NN architecture and the back propagation [68], [69] or evolutionary techniques are used to adjust the weights of the NN model. In general, NN based models are also used for data classification [70] [71].

Figure 5-13: Classification accuracies of the SVM classification model plotted against γ values (model parameter) for a probability estimate of one.
5.1.1.1. GA based NN Model

The NN model developed in this dissertation has 4 neurons in the input layer, 4 neurons in the hidden layer, and one neuron in the output layer, implying that there are 4 inputs and one output. Inputs to this NN model are a training function, $\gamma$ value, probability estimate, and original classification accuracy from SVM model. Total instances (data samples) used for training the neural network model are 1564. This newly developed training data represent all parameters required for training and testing the NN model - four inputs and one output. Given multiple sets of parameters used for training and testing the SVM model, and the available CPU time, NN model can predict the classification accuracy and vice versa.

The NN model is activated using a sigmoid function, and this model is then tested using the GA Tool. In the GA tool evaluation, number of generations performed is 100, and the obtained mean fitness value is closer to the best fitness value after first 50
generations. Either instantaneous energy data or encoded data can be used for estimating the optimum set of required SVM model parameters and the selected set of parameters are then used for classifying customers in real-time. Figure 5-15 represents the chromosome of the NN model. The chromosome used in this NN model has 20 weights and 5 thresholds. T1, T2, T3, T4, and T5 are the thresholds of the neurons 1, 2, 3, 4, 5 respectively. Here, the chromosome representing the NN model has 25 genes. Using the GA tool, the best chromosomes will be evolved and the best chromosome represents NN weights with minimum classification error [72]. Among the total population of chromosomes, a chromosome with high fitness has a greater probability of getting selected as a parent; crossover and mutation are operated on selected chromosomes.

Figure 5-16 illustrates the procedure implemented for generating training data and operation of the resultant NN model. The significant achievement of this model is that the GA convergence graph for the number of generations vs. error has converged. The best fitness error is $3.496 \times 10^{-6}$ and the mean error is 0.0688 as shown in Figure 5-17. These values are promising, as the error is very small and the NN model is converging at about 50 generations. The chromosomes applied in this model are able to train and estimate the classification accuracy of the energy consumption data. One stage NN model did not yield superior performance in the estimation of data classification accuracy. In this model, all the input values are given to this hierarchical NN model, wherein, these data samples will go through several levels of the NN and the final result would be an enhanced estimation of data classification accuracy. Advantages of using the NN model - this model save time in selecting the parameters for SVM model, which would be significant in real time analysis.
Figure 5-15: Structure of the chromosome used in the developed NN model.

Figure 5-16: Procedure for generating the data required to train the NN model and estimate SVM model parameters using the NN model.
5.2 Rule-Based Classification Model

The second stage of customer energy consumption profile evaluation is initiated by inputting the encoded energy consumption data to the rule engine model. Recently, rule engines are being used in several applications like classification, pattern recognition, fault analysis, data mining, search engines, strategy making, etc. [73]–[80]. Customers are classified into three classes based on irregularities in energy consumption. These rules are framed such that most forms of illegal consumption can be identified. Generalized rules for classification are,

For being classified under class-D: in any sample/instance of encoded inputs,

- Two or more inputs should be zero and/or
- Six or more inputs are an integer between 0 and 6.
For being under class-S: in any sample/instance,

- Not more than one input can be ‘0’ and/or
- Not more than four inputs are an integer between 0 and 6.

For being under class-I: in any sample/instance,

- Not more than three inputs are numbers between 1 and 6.

Based on these generalized rules, some specific rules are developed that collectively form the rule engine. Here, the total number of rules developed is 71. Of these, 47 are intended for identifying illegal consumers, 23 for suspicious consumers, and one for grouping genuine consumers. In addition, rules are framed such that leverage has been given for illegal consumers over genuine customers, to ensure any genuine consumer will not be categorized as illegal consumer. For illustration, some of the rules implemented/used are presented. The number of rules is limited as a compromise between the model accuracy, CPU time consumed and complexity of rules (it has been made sure that the rules can be debugged and modified if necessary). Figure 5-18 illustrates some of the rules used in the algorithm. If the distribution losses exceed 4%, the rule engine algorithm will be triggered again. If not, the algorithm may be terminated until reception of the next cycle of data.

Here, ‘count’ represents the number of irregularities of a particular customer; ‘count’ is the sum of values other than ‘7’ (i.e. ‘0, 1, 2, 3, 4, 5, 6’) in each customer’s encoded consumption profile. In this data, numbers ‘0’ through ‘6’ in any customer’s encoded energy consumption has been considered as probable irregularities and these numbers are added individually. Therefore, now we have a list of irregularities, i.e. sums of numbers 1 to 6 (extracted from individual consumer’s encoded energy consumption),
in a 24 hour period. Figure 5-19 illustrates the sum of irregularities of all consumers in a single day. Different colored lines in this figure illustrate different irregularities for each customer. Figure 5-20 aids in understanding the occurrence of irregularities in a customer’s energy consumption profile. Figure 5-20 illustrates the performance of the proposed algorithm. It displays the customer classes predicted plotted against actual classes. Un-overlapped dots in the figure represent improperly classified customers.

<table>
<thead>
<tr>
<th>For class-D;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. if count(p,1) &gt; 2</td>
</tr>
<tr>
<td>2. elseif count(p,2) &gt; 3</td>
</tr>
<tr>
<td>3. elseif count(p,7) &gt; 6</td>
</tr>
<tr>
<td>4. elseif count(p,1) &gt; 0 &amp;&amp; count(p,1) &lt; 2 &amp;&amp; count(p,2) &gt; 1</td>
</tr>
<tr>
<td>5. elseif count(p,2) &gt; 1 &amp;&amp; count(p,5) &gt; 3</td>
</tr>
<tr>
<td>6. elseif count(p,1) &gt; 1 &amp;&amp; count(p,2) &gt; 1 &amp;&amp; count(p,3) &gt; 1 &amp;&amp; count(p,5) &gt; 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>For class-S;</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. elseif count(p,1) &gt; 0 &amp;&amp; count(p,1) &lt; 2</td>
</tr>
<tr>
<td>8. elseif count(p,5) &gt; 1 &amp;&amp; count(p,6) &gt; 1</td>
</tr>
<tr>
<td>9. elseif count(p,3) &gt; 1 &amp;&amp; count(p,4) &gt; 1 &amp;&amp; count(p,5) &gt; 1 &amp;&amp; count(p,6) &gt; 1</td>
</tr>
</tbody>
</table>

Figure 5-18: Synopsis of the rules used to develop the rule engine based classification algorithm.

The proposed rule engine algorithm has been applied on the encoded data as it requires less time for classification, lesser number of rules, and easy to analyze and alter, compared to regular energy consumption data. The proposed algorithm has consumed
0.4992 seconds (0.156 sec for data processing, 0.1092 sec for classification) and classified 802 consumers correctly out of 882. The classification accuracy is about 91%.

Figure 5-19: Sum of individual irregularities in energy consumption of 450 customers’ over a single day.

Figure 5-20: Classes classified by the rule engine classification algorithm plotted against actual classes of consumers.
5.3 Multi-Level Classification Model

At this stage of classification, every customer is now assigned a particular class by both the rule engine and SVM classification algorithms. These classes are compared to see if a particular customer is classified into same or different classes by the two algorithms as shown in Figure 5-21 (a). As described earlier, energy consumption profile of every customer is evaluated individually by both the rule engine and SVM classification algorithms. If both algorithms yield the same result for a customer, then that customer’s class will be finalized, and appropriate action will be executed. If the result from both these algorithms is different, then those particular energy consumption patterns are sent for reclassification by Neural Network Pattern Recognition (NNPR) tool. Entire classification process can be terminated at a point when the overall distribution losses are less than 4% excluding the portion consumed by classified illegal consumers. Figure 5-21 (b) illustrates the actions proposed to be performed on customers based on their classes. For one attempt of operation of multilevel classification algorithm, proposed algorithm has consumed 0.365 seconds (0.156 sec for encoding, 0.1092 sec for rule engine classification and 0.1 sec for SVM model) for classifying 882 customers.
Figure 5-21 (a): Multilevel classification algorithm for identifying illegal consumers.

Figure 5-21 (b): Actions proposed to be taken on illegal consumers.
5.4 NNPR Tool

This NNPR classification model has been used to classify those customers who were assigned different classes by Rule Engine and SVM model. The required classification model has been developed from NNPR tool, a MATLAB toolbox [81]. This model is used to evaluate energy consumption of certain customers sent for reclassification by the multilevel classification algorithm in Figure 5-21 (a). This tool allows a user to load energy consumption data and output classes for training and testing purposes.

In this tool, a two-layer feed-forward network, with sigmoid hidden and output neurons, classifies vectors arbitrarily. The network will be trained with scaled conjugate gradient back propagation. Figure 5-22 displays the conceptual representation of NNPR model. Total number of neurons in the hidden layer is 110. Figure 5-23 displays the ROC curves for the classification. For illustration, in this section, NNPR model has been used to classify encoded energy consumption (patterns of irregularities) of 882 customers. Overall classification accuracy of NNPR model for classifying the encoded energy consumption data is 85.5%. Figure 5-24 displays the classes predicted vs. actual classes mapped over customer IDs. Non-overlapping lines in the figure represent incorrectly classified customers. During real-time operation, the NNPR classification tool is used only for energy consumption data of customers that are classified differently by SVM model and Rule Engine, and as a result, this classification model may have to be applied on about 15% of overall data (for patterns of about 15% of all customers).
Figure 5-22: Neural network structure of the NNPR based classification model.

Figure 5-23: ROC curves generated by the NNPR classification model.
Figure 5-24: Classes predicted by the NNPR based classification model plotted against the actual classes of the testing data.

5.5 Summary

To summarize Chapter 5, intelligent data classification algorithms for classifying illegal consumers based on their energy consumption are elucidated. These methods are designed to detect illegal consumers in a smart grid environment where utilities have access to customers’ real-time energy consumption data. SVM model, GA based NN model, Rule Engine algorithms, and NNPR tools have been developed to classify and group illegal consumers. The process, order and priority in which these tools are modeled and used are explained.
Chapter 6

Identifying Illegal Consumers of Electricity using High Performance, Intelligent Classification

Over the past decade, several power engineering, monitoring, control and optimization applications, power systems analysis and simulations are performed on HPC to improve their performance in real-time. Recently, due to the advent of smart grid, algorithms used to perform power engineering tasks have become further complex and computationally intensive. HPC aids in the real-time analysis, visualization, and intelligent management of the grid [82]. In real-time, operation of data processing and encoding techniques should be a continuous process; operation of the SVM model and Rule Engine will be operated accordingly. Broadly, the classification procedure falls into three major categories - SVM model, Rule Engine, and multilevel classification. In the Multi-level classification algorithm implementation, classification results from SVM model and Rule Engine are analyzed, prioritized and finalized. SVM, Rule Engine and the hybrid multilevel classification models proposed in the dissertation, classify a few hundred of customers in a few seconds. However, when these algorithms are operated on thousands of consumers, millions of data points have to be analyzed in real-time.
Therefore, these algorithms presented in chapters 4 and 5 are modified to explore advantages of HPC. This chapter explains the procedure implemented to parallelize identification of illegal consumers from the encoded data.

### 6.1 High Performance Computing

The advancements in science and technology led to a rapid increase in demand for computational power. HPC helps scientists and engineers solve several complex science and engineering problems that require high bandwidth, low latency networking, and very high computational capabilities. The term HPC can be applied to any system that is capable of performing 1012 to 1015 floating-point operations per second i.e. teraflops to petaflops [83]. The key industrial and societal applications of HPC in design and manufacturing include automotive and aerospace, services and utilities including energy, digital media, and financial services, healthcare and medicine, efficient transportation, and emergency response [84]. The advent of smart grid makes power systems analysis complex and computationally intense and offers an effective platform to exploit the benefits of HPC to the full extent.

The trends in the HPC can be broadly classified into Algorithmic trends and Computational trends. Algorithmic trends include decomposing, parallelizing, or implementing a specific problem using some novel techniques, on the other hand, computational trends relate to the hardware and software technologies to solve problems of higher dimensions [85].
6.1.1. Algorithmic Trends

The algorithmic trends are inspired from the strategy ‘divide and conquer’. The problem is decomposed, and parallelization techniques are employed to exploit the benefits of HPC. In problem decomposition, the problem in hand is decomposed into several independent tasks and these tasks are implemented concurrently on different machines to speed up the process. For example, suppose the problem is to find summation of numbers 1 through 10. In this case, the problem can be decomposed into two sub problems with one adding numbers 1 to 5 and the other adding numbers 6 to 10. These two processes are independent and can be run simultaneously on different CPUs and finally the results from both processes are added to yield the final result. However, it is not always possible to decompose a problem.

The parallelization techniques can be broadly classified into data parallelism and task parallelism. In task parallelism, the given task is subdivided into multiple tasks, which can be mapped to and solved on multiple platforms with varying computational power. In task parallelism, different calculations are performed on either different sets or same set of data. In case of data parallelism, the same operation is performed on different sets of data [86]. The problems in power system analysis rely heavily on matrix computations and independent sub-problems. Hence, both data and task parallelism are feasible for power system analysis.
6.1.2. Computational Trends

The motive for parallel computing started from a basic idea; $n$ computers operating simultaneously can achieve the result $n$ times faster (ideally). In practice, it will not be $n$ times faster because of certain factors and limitations. The parallel computing can be generally classified into Multi-core computing, Symmetric Multiprocessing, Cluster Computing, Grid Computing, and Graphic Processing Units (GPUs) etc.

A Multi-core processor is a computational unit with multiple processing units called cores. Each processing unit reads and executes program instructions simultaneously, thereby speeding up the process [87]. All the cores are typically integrated onto a single chip. The information flow between the cores can be achieved by either passing a message or sharing cache memory. As Moore’s law begins to fade, multicore technology draws the attention of manufacturers as well as customers.

Systematic Multiprocessing (SMP) involves two or more processors that are connected to single shared memory via system bus or crossbar switch. The task can be moved efficiently and easily between the processors to balance the work load. The bandwidths of either bus or crossbar switch and power consumption of various processors are the bottleneck in the scalability of SMP. Serious challenges remain to be answered with this architecture as it needs two modes of programming including one for the CPU itself and another for interconnect between CPUs.

Whereas, Cluster Computing contains a group of several loosely coupled computers and are viewed as a single system [88]. The cluster contains a number of individual computers connected over a high speed LAN. Cluster computing is a cost
effective solution to high performance computing as cluster comprises multiple personal computers.

Grid Computing is similar to Cluster Computing. The major difference between a cluster and a grid is that, a cluster is homogeneous (i.e. all computers have similar hardware and operating system) while a grid is heterogeneous (i.e. the computers connected in the grid have different architecture and a different operating system). Another difference is that, usually the clusters are distributed over a building or an entity; on the other hand the grid is sometimes connected via Wide Area Network (WAN) or Metropolitan Area Network (MAN) [89].

Recently, GPU computing has been gaining attention in the HPC market. GPUs are primarily built to accelerate the process of building images in a frame buffer for display. Highly parallel structure of GPU makes it more efficient compared to general-purpose CPU. GPUs can process only independent fragments, but multiple fragments in parallel. This very ability of GPUs makes it apt for power system analysis [90]–[95]. Ideal GPU applications have large data sets and minimal dependency between data elements. In addition, multiple GPUs can be employed simultaneously making the already parallel process even more parallel and thereby reducing the computational time.

### 6.2 Data Loading and Encoding

The process of loading and encoding the energy consumption data has been parallelized. As presented in classification methodology of chapter 5, classification algorithms require 4 data files (training input, training output, testing input, and testing
output); this study examines two distinct methods for loading data. In the first, the data is loaded in a typical and sequential fashion. In the second, the data is loaded in a task parallel manner, with each file being loaded simultaneously by separate threads.

Implementation of the encoding technique used to simplify the energy consumption data and map instantaneous energy consumption patterns into consumption irregularities is explained in chapter 4. However, this process has been further modified such that portions of the entire process are executed in parallel; i.e. loading the data, calculations or other modifications are executed in parallel. In the entire encoding process, calculation of the following parameters and processes are parallelized:

- Calculation of Average consumption values
- Calculation of MF values
- Calculation of AOCs - rows are parallelized
- Calculation of CPC - rows are parallelized
- Second stage of encoding has been parallelized such that some tasks between multiple time zones are executed in parallel using data parallelism and task parallelism.

In spite of the situation that some of the calculations depend on the historic data (initial three days of the month), there has not been much data dependency issues. Results from serial execution and parallel execution are presented in the results section.
6.3 Identification of Illegal Customers

Illegal consumers on the grid are detected using the classification algorithms illustrated in chapter 5. This section of the dissertation explains the process implemented in parallelizing both classification algorithms and the process of loading the data to the algorithms. Identification of illegal customers has been done in three stages: SVM Model, rule engine, and Multi-level classification model.

6.3.1. SVM Based Classification Model

In this dissertation, LibSVM has been used to develop the required SVM based classification model. SVM based classification model used for identifying illegal consumers is illustrated in chapter 5. Training and testing the SVM model requires us to select a kernel function for training the model, parameters involved in the training function, probability estimates, and the cost function. After modeling the required SVM based classification model, the customer energy consumption profile evaluation has been initiated by inputting the processed/encoded data to the SVM model. Parameters used for SVM classification model are:

- Kernel function: Sigmoid, C=1, γ = 0.3.
- The testing data has 20,000 samples. The training data was kept at 442.
- The code written was parallelized using the Microsoft Concurrency Runtime.
- The SVM was parallelized in the case of Data/Hybrid parallelism using...
OpenMP using 8 cores.

The SVM algorithm was trained offline using 5-fold validation.

Rules for classification are similar to the rules used by classification models for encoded data in chapter 5. Based on these rules, customers classified as class-D are potential illegal consumers, class-S are suspicious consumers, and class-I are genuine consumers. Figure 6-1 illustrates classification accuracy of the SVM model for different combinations of parameters and the classification accuracy of the SVM model ranged between 76 and 92%. Resultant CPU times for classification of 20,000 customers both in serial execution as well as different types of parallel processing is presented in the results section.

![Classification Accuracy](image)

**Figure 6-1:** Classification accuracy of the SVM based classification model for different attempts of classification i.e. for different sets of model parameters.
6.3.2. Rule-Based Classification Model

The next stage of customer energy consumption profile evaluation is initiated by inputting the encoded energy consumption data to the Rule Engine model. Rules used for classification are similar to the rules used in Rule Engine model presented in chapter 5. Figure 6-2 illustrates the sum of irregularities of all consumers in a 24-hour day period. Different color lines in this figure represent different irregularities (0, 1, 2, 3, 4, 5, and 6) for each customer. Figure 6-2 aids in understanding the occurrence of irregularities in a customer’s energy consumption profile. Figure 6-3 illustrates the performance of the proposed algorithm. It plots the customer classes predicted against actual classes. Black dots in the figure represent improperly classified customers or error in classification. Classification accuracy of the Rule Engine based classification model is about 92%.

Figure 6-2: Sum of individual irregularities in energy consumption of 450 customers’ over a single day.
Figure 6-3: Classes predicted by the rule based classification algorithm plotted against actual consumer classes.

In the rule engine based classification model, the following procedures have been parallelized:

- Irregularities (inputs between ‘0’ and ‘6’) in encoded energy consumption readings of multiple customers are counted in parallel at the same time.
- Energy consumption patterns of multiple customers are evaluated in parallel at the same time.

**6.3.3. Multi-Level Classification Model**

At the end of previous stage, every customer will be assigned a particular class by both the Rule Engine and SVM classification algorithms. These classes are compared to
see if a customer is classified into same or different classes by the two algorithms. If both algorithms yield the same class for a customer, then that customer’s class will be finalized, and appropriate action will be executed. If the result from both these algorithms is different, then those particular energy consumption patterns are sent for reclassification by NNPR tool as illustrated in chapter 5. Entire classification algorithm can be terminated at a point when the overall distribution losses are less than 4%. The entire classification algorithm has been operated on energy consumption patterns of 20,000 customers. Figure 6-4 illustrates the task parallel implementation of the algorithm. Figure 6-5 illustrates data parallel implementation on the classification algorithm. These two algorithms were implemented and the results are presented.

Figure 6-6 illustrates an alternative parallelization approach to the classification algorithm. This alternative approach may be implemented as follows: first, customer energy consumption is encoded; then, the SVM and Rule Engine classification algorithms are applied; results from these algorithms are compared; if results are same, then the customer’s class is finalized, if they are not equal, then the energy consumption profile has to be re-evaluated and reclassified using NNPR tool; this entire procedure is executed as a single unit for each customer - exclusively. In a similar way, this procedure is executed for all customers in parallel explicitly, such that multiple customers energy consumption data is evaluated which improves the overall CPU time.

As discussed in the earlier sections of the chapter, data loading, encoding and classification algorithms have been executed in a serial mode as well as in different types of parallelizing techniques. Table 6.1 illustrates the combinations of serial and parallel techniques implemented in different stages of the algorithm.
Figure 6-4: Task Parallel implementation of the entire multi-level classification algorithm.
Figure 6-5: Data Parallel implementation of the entire multi-level classification algorithm.

Figure 6-6: Alternative process for parallelizing the proposed multi-level classification algorithm.
Table 6.1 Stages involved in execution of the entire algorithm.

<table>
<thead>
<tr>
<th></th>
<th>Serial execution</th>
<th>Task par.</th>
<th>Data par.</th>
<th>Full parallel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Loading</td>
<td>Serial</td>
<td>Parallelize</td>
<td>Serial</td>
<td>Task par.</td>
</tr>
<tr>
<td>Encoding</td>
<td>Serial</td>
<td>Parallelize</td>
<td>Serial</td>
<td>Task par.</td>
</tr>
<tr>
<td>Averages &amp; AOC</td>
<td>Serial</td>
<td>Parallelize</td>
<td>Parallelize</td>
<td>Task par.</td>
</tr>
<tr>
<td>Calculating MF &amp; summing columns</td>
<td>Serial</td>
<td>Parallelize</td>
<td>Parallelize</td>
<td>Parallelize</td>
</tr>
<tr>
<td>Rule Engine</td>
<td>Serial</td>
<td>Serial</td>
<td>Parallelize</td>
<td>Parallelize</td>
</tr>
<tr>
<td>SVM model</td>
<td>Serial</td>
<td>Parallelize</td>
<td>Parallelize</td>
<td>Parallelize</td>
</tr>
<tr>
<td>Comparison</td>
<td>Serial</td>
<td>Serial</td>
<td>Serial</td>
<td>Serial</td>
</tr>
</tbody>
</table>

6.4 Results and Discussion

Combined classification accuracy all classification models is about 89%. As illustrated in earlier sections, overall algorithm has been parallelized in several combinations at different stages. This section of the dissertation illustrates the results obtained in different stages of the parallelized algorithm.

In the following tables, the following conventions are used:

- **LD** – Time consumed for *loading data*
- **EN** – Time consumed for *data encoding*
- **CAL** – Time consumed for *calculation*
- **RE** – Time consumed for *Rule Engine based classification model*
- **SVM** – Time consumed for *SVM based classification model*

Table 6.2 illustrates the time taken for each of the processes explicitly in ten random attempts and the average CPU time. It has been observed that, in serial execution of the algorithm, average time for executing the overall algorithm is about 17 seconds;
and most of this time was consumed for loading the data.

Table 6.2 CPU time for serial execution of the algorithm.

<table>
<thead>
<tr>
<th>Time</th>
<th>LD</th>
<th>EN</th>
<th>CAL</th>
<th>SVM</th>
<th>RE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.96</td>
<td>0.19</td>
<td>0.13</td>
<td>0.27</td>
<td>0.04</td>
<td>17.59</td>
<td></td>
</tr>
<tr>
<td>16.40</td>
<td>0.17</td>
<td>0.14</td>
<td>0.33</td>
<td>0.05</td>
<td>17.09</td>
<td></td>
</tr>
<tr>
<td>16.57</td>
<td>0.17</td>
<td>0.14</td>
<td>0.29</td>
<td>0.05</td>
<td>17.22</td>
<td></td>
</tr>
<tr>
<td>16.61</td>
<td>0.17</td>
<td>0.14</td>
<td>0.26</td>
<td>0.05</td>
<td>17.23</td>
<td></td>
</tr>
<tr>
<td>16.73</td>
<td>0.16</td>
<td>0.13</td>
<td>0.34</td>
<td>0.05</td>
<td>17.41</td>
<td></td>
</tr>
<tr>
<td>16.54</td>
<td>0.17</td>
<td>0.14</td>
<td>0.31</td>
<td>0.05</td>
<td>17.21</td>
<td></td>
</tr>
<tr>
<td>16.77</td>
<td>0.17</td>
<td>0.13</td>
<td>0.29</td>
<td>0.05</td>
<td>17.41</td>
<td></td>
</tr>
<tr>
<td>16.61</td>
<td>0.17</td>
<td>0.13</td>
<td>0.24</td>
<td>0.05</td>
<td>17.2</td>
<td></td>
</tr>
<tr>
<td>16.57</td>
<td>0.17</td>
<td>0.14</td>
<td>0.26</td>
<td>0.05</td>
<td>17.19</td>
<td></td>
</tr>
<tr>
<td>16.53</td>
<td>0.17</td>
<td>0.13</td>
<td>0.26</td>
<td>0.05</td>
<td>17.14</td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>16.63</td>
<td>0.17</td>
<td>0.13</td>
<td>0.28</td>
<td>0.05</td>
<td>17.26</td>
</tr>
</tbody>
</table>

In task parallel implementation of the overall algorithm, time consumed for most portions of the algorithm is reduced; however, time taken for loading data has been significantly reduced. Therefore, the average time for the entire algorithm is about 9 seconds as illustrated in table 6.3.

Table 6.3 CPU time for Task parallel execution of the algorithm.

<table>
<thead>
<tr>
<th>Time</th>
<th>LD</th>
<th>EN</th>
<th>CAL</th>
<th>SVM</th>
<th>RE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.08</td>
<td>0.11</td>
<td>0.07</td>
<td>0.64</td>
<td>0.04</td>
<td>9.94</td>
<td></td>
</tr>
<tr>
<td>9.06</td>
<td>0.11</td>
<td>0.09</td>
<td>0.63</td>
<td>0.05</td>
<td>9.94</td>
<td></td>
</tr>
<tr>
<td>8.85</td>
<td>0.10</td>
<td>0.08</td>
<td>0.64</td>
<td>0.05</td>
<td>9.72</td>
<td></td>
</tr>
<tr>
<td>8.86</td>
<td>0.10</td>
<td>0.09</td>
<td>0.64</td>
<td>0.05</td>
<td>9.74</td>
<td></td>
</tr>
<tr>
<td>8.91</td>
<td>0.11</td>
<td>0.08</td>
<td>0.64</td>
<td>0.05</td>
<td>9.79</td>
<td></td>
</tr>
<tr>
<td>8.89</td>
<td>0.10</td>
<td>0.08</td>
<td>0.63</td>
<td>0.05</td>
<td>9.75</td>
<td></td>
</tr>
<tr>
<td>8.94</td>
<td>0.10</td>
<td>0.08</td>
<td>0.63</td>
<td>0.05</td>
<td>9.8</td>
<td></td>
</tr>
<tr>
<td>8.82</td>
<td>0.10</td>
<td>0.08</td>
<td>0.67</td>
<td>0.05</td>
<td>9.72</td>
<td></td>
</tr>
<tr>
<td>8.94</td>
<td>0.10</td>
<td>0.08</td>
<td>0.64</td>
<td>0.05</td>
<td>9.81</td>
<td></td>
</tr>
<tr>
<td>8.86</td>
<td>0.10</td>
<td>0.08</td>
<td>0.63</td>
<td>0.05</td>
<td>9.72</td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>8.92</td>
<td>0.10</td>
<td>0.08</td>
<td>0.64</td>
<td>0.05</td>
<td>9.79</td>
</tr>
</tbody>
</table>

In the execution of data parallelism, time consumed for every portion of the
algorithm is slightly less except for data loading. Table 6.4 displays the execution times in data parallel implementation. Therefore, the average time for the entire algorithm remained very close to the serial execution.

Table 6.4 CPU time for Data parallel execution of the algorithm.

<table>
<thead>
<tr>
<th></th>
<th>LD</th>
<th>EN</th>
<th>CAL</th>
<th>SVM</th>
<th>RE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempts</td>
<td>17.17</td>
<td>0.23</td>
<td>0.09</td>
<td>0.53</td>
<td>0.03</td>
<td>18.05</td>
</tr>
<tr>
<td></td>
<td>15.24</td>
<td>0.10</td>
<td>0.10</td>
<td>0.28</td>
<td>0.03</td>
<td>15.75</td>
</tr>
<tr>
<td></td>
<td>15.32</td>
<td>0.10</td>
<td>0.11</td>
<td>0.23</td>
<td>0.03</td>
<td>15.79</td>
</tr>
<tr>
<td></td>
<td>15.28</td>
<td>0.11</td>
<td>0.12</td>
<td>0.27</td>
<td>0.03</td>
<td>15.81</td>
</tr>
<tr>
<td></td>
<td>15.32</td>
<td>0.10</td>
<td>0.10</td>
<td>0.31</td>
<td>0.03</td>
<td>15.86</td>
</tr>
<tr>
<td></td>
<td>15.29</td>
<td>0.10</td>
<td>0.10</td>
<td>0.25</td>
<td>0.03</td>
<td>15.77</td>
</tr>
<tr>
<td></td>
<td>15.23</td>
<td>0.10</td>
<td>0.11</td>
<td>0.27</td>
<td>0.03</td>
<td>15.74</td>
</tr>
<tr>
<td></td>
<td>15.27</td>
<td>0.13</td>
<td>0.11</td>
<td>0.28</td>
<td>0.03</td>
<td>15.82</td>
</tr>
<tr>
<td></td>
<td>15.40</td>
<td>0.10</td>
<td>0.11</td>
<td>0.25</td>
<td>0.03</td>
<td>15.89</td>
</tr>
<tr>
<td></td>
<td>15.25</td>
<td>0.10</td>
<td>0.10</td>
<td>0.27</td>
<td>0.03</td>
<td>15.75</td>
</tr>
<tr>
<td>Avg.</td>
<td>15.48</td>
<td>0.12</td>
<td>0.10</td>
<td>0.29</td>
<td>0.03</td>
<td>16.02</td>
</tr>
</tbody>
</table>

Combined parallel algorithm has been executed such that every parallelizable portion of the overall algorithm is parallelized. In this combined parallel execution, some portions of the algorithm have been executed in task parallelism, and the rest in data parallelism such that the average execution time for the entire algorithm is reduced to half the time consumed for serial execution. Results are displayed in table 6.5.

Table 6.5 CPU time for combined parallel execution of the algorithm.

<table>
<thead>
<tr>
<th></th>
<th>LD</th>
<th>EN</th>
<th>CAL</th>
<th>SVM</th>
<th>RE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempts</td>
<td>9.08</td>
<td>0.11</td>
<td>0.07</td>
<td>0.22</td>
<td>0.04</td>
<td>9.52</td>
</tr>
<tr>
<td></td>
<td>9.06</td>
<td>0.11</td>
<td>0.09</td>
<td>0.28</td>
<td>0.05</td>
<td>9.59</td>
</tr>
<tr>
<td></td>
<td>8.85</td>
<td>0.10</td>
<td>0.08</td>
<td>0.26</td>
<td>0.05</td>
<td>9.34</td>
</tr>
<tr>
<td></td>
<td>8.86</td>
<td>0.10</td>
<td>0.09</td>
<td>0.40</td>
<td>0.05</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>8.91</td>
<td>0.11</td>
<td>0.08</td>
<td>0.25</td>
<td>0.05</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>8.89</td>
<td>0.10</td>
<td>0.08</td>
<td>0.94</td>
<td>0.05</td>
<td>10.06</td>
</tr>
<tr>
<td></td>
<td>8.94</td>
<td>0.10</td>
<td>0.08</td>
<td>0.25</td>
<td>0.05</td>
<td>9.42</td>
</tr>
<tr>
<td></td>
<td>8.82</td>
<td>0.10</td>
<td>0.08</td>
<td>0.27</td>
<td>0.05</td>
<td>9.32</td>
</tr>
</tbody>
</table>
Table 6.6 shows time consumed for serial and parallelized execution for portions of the algorithm explicitly.

Table 6.6 Average times for ten attempts of execution.

<table>
<thead>
<tr>
<th></th>
<th>LD</th>
<th>EN</th>
<th>CAL</th>
<th>SVM</th>
<th>RE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial</td>
<td>16.63</td>
<td>0.17</td>
<td>0.13</td>
<td>0.28</td>
<td>0.05</td>
<td>17.26</td>
</tr>
<tr>
<td>Data Parallel</td>
<td>15.48</td>
<td>0.12</td>
<td>0.10</td>
<td>0.29</td>
<td>0.03</td>
<td>16.02</td>
</tr>
<tr>
<td>Task Parallel</td>
<td>8.92</td>
<td>0.10</td>
<td>0.08</td>
<td>0.64</td>
<td>0.05</td>
<td>9.79</td>
</tr>
<tr>
<td>Combined Par.</td>
<td>8.92</td>
<td>0.10</td>
<td>0.08</td>
<td>0.33</td>
<td>0.05</td>
<td>9.48</td>
</tr>
</tbody>
</table>

Table 6.7 shows the speedup achieved by implementing parallelism to the entire algorithm.

Table 6.7 Average Speedups achieved.

<table>
<thead>
<tr>
<th></th>
<th>LD</th>
<th>EN</th>
<th>CAL</th>
<th>SVM</th>
<th>RE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Data Parallel</td>
<td>1.07</td>
<td>1.42</td>
<td>1.30</td>
<td>0.97</td>
<td>1.67</td>
<td>1.08</td>
</tr>
<tr>
<td>Task Parallel</td>
<td>1.86</td>
<td>1.70</td>
<td>1.63</td>
<td>0.44</td>
<td>0.98</td>
<td>1.76</td>
</tr>
<tr>
<td>Combined Par.</td>
<td>1.86</td>
<td>1.70</td>
<td>1.62</td>
<td>0.85</td>
<td>0.98</td>
<td>1.82</td>
</tr>
</tbody>
</table>

From the results, it has been observed that the proposed algorithm would yield better results (less computational time) when applied on a much bigger data set, which is the case in real-world implementation. However, based on the speedup in the results, data loading may be done using task parallelism, encoding the data may also be done using task parallelism, running the Rule Engine may be done using data parallelism.
6.5 Summary

To summarize Chapter 6, the importance of implementation of algorithms for detection of illegal consumers on HPC is discussed. Data loading, encoding presented in chapter 4 and classification algorithms presented in chapter 5 are modified to explore the advantages of parallel programming and processing. The newly developed algorithms supporting Task Parallelism and Data Parallelism approaches are illustrated in this chapter. The obtained results in terms of improvement in computational time are presented.
Chapter 7

Impact of Time-Based Pricing and Distributed Generation on Illegal Consumption of Electricity

A smart grid is an intelligent upgrade to the existing electric infrastructure. The smart grid offers better optimization and utilization of existing assets, self-healing capability, improved monitoring and control of bulk generation sources, DG sources, faster relays and enhanced protection with advanced relays, re-closures, and other intelligent infrastructure. On the other hand, customer options offered by smart grid include complete monitoring and control over household energy consumption at the customer end, and flexibility in operation of appliances. These options and monitoring tools will help customers to reduce energy consumption and enhance the living experience at home. Unfortunately, these customer options may affect the performance of tools designed for detecting and controlling illegal consumers. Therefore, these detection tools should consider the impact of features offered at the customer end. Those features to be considered are RTP, DG, PHEVs, and Demand Side Management (DSM).
7.1 Dynamic or Time-based Pricing (TBP)

In dynamic pricing or TBP, unlike a flat rate tariff, the utility charges its customers at a tariff that depends on the time, when that power is supplied. Generally, in TBP, the tariff would be high when the overall demand or load on the grid is high and the tariff would be significantly less when the overall demand or load is less. TBP is implemented by utilities both in regulated as well as market based environments. However, TBP may not be implemented if the overall electricity demand is not high. In a regulated power market, TBP will be developed based on the cost involved in operation and investment costs on a long run. Some important types of TBP include Critical Peak Pricing (CPP), Time-Of-Use pricing (TOU) and Real-Time Pricing (RTP).

7.1.1. CPP

CPP tariff scheme is a recently introduced type of TBP, where peak period tariffs are more targeted. In CPP, high tariff zones or on-peak period times are limited to a few days per year. These days are determined based on the predictions that energy consumption (overall demand) would be very high. The goal of CPP is to shift customer energy consumption from on-peak hours to off-peak hours. As an example, utilities may impose a CPP tariff window between 2:00 PM and 7:00 PM and for up to 6 to 15 days per year. Customers accepted into a CPP tariff scheme will have a lower tariff during the off-peak period as a compensation for paying higher tariff during the peak period. During
on-peak periods, tariff typically range between 400% and 700% of the off-peak electricity rate. The difference in tariff acts as an incentive for customers to reduce their consumption during critical on-peak periods which enhances the utility’s day to day operations such as demand shaving, management and shifting.

7.1.2. TOU Pricing

As illustrated in chapter 3, Advanced Metering Infrastructure (AMI) will be able to monitor and collect the instantaneous energy consumption information of individual customers with time-stamp of consumption. Making use of AMI, TOU pricing plans have been developed where a residential customer pays higher tariff for energy utilized during the on-peak period compared to the off-peak period. This plan has also been developed to discourage customers from using power during the on-peak period. As an example, Baltimore Gas and Electric (BGE) offers its customers a program with five different plans based on time at which electricity is consumed. The tariff structure is as follows: cost is 10.32 ¢/kWh from 7:00 - 10:00 AM and 8:00 – 11:00 PM, 9.305 ¢/kWh between 11:00 PM and 7:00 AM, and 15 ¢/kWh between 10:00 AM and 8:00 PM.

7.1.3. RTP

In the RTP model, the tariff of the electricity consumed by a customer changes in real-time based on the supply and demand. Therefore, the tariff during an on-peak period
would be high on a few extremely hot summer days. In general, such days are same as the days that CPP is imposed. Therefore, RTP may not be helpful in controlling, shifting or conserving energy demand all through the year with a few exceptions (where demand significantly outweighs supply) [82], [96]-[102]. Figure 7-1 illustrates RTP over an entire day on May 02, 2012. It can be observed that the tariff on power consumed around 2:00 PM is about 5¢/kWh and less than 2¢/kWh around 2:00 AM. On-peak period tariff is about 250% of tariff during the off-peak period.

Figure 7-1: Day Ahead Pricing (in $/kWh) used for billing RTP customers [103].
7.2 Impact of TBP

Independent of the type of TBP implemented, customers will have to pay higher tariff during the on-peak period compared to rest of the time. As illustrated in chapter 2, several factors affect illegal consumers pilfering of electricity. However, TBP could be a significant factor that influences illegal consumers to steal electricity during on-peak periods. Illegal consumption of electricity during the on-peak period may impact the primary objective of TBP, i.e. reducing the energy consumption. Because of illegal consumption, energy consumption may increase instead of decreasing or at least compensate for the quantity of energy conserved by genuine consumers. As a result, overall demand on the grid doesn’t decrease much during the on-peak period. In addition, quantity of illegally consumed electricity during the on-peak periods may be much higher than illegally consumed electricity during the off-peak periods. Therefore, quantity of illegally consumed energy may be directly proportional to the tariff at which electricity is supplied to the customers (prominently during the on-peak period). In addition to the demand optimization and management, enhanced power quality, and other day-to-day operations performed by utilities and ISOs, detection of illegal consumers should also be prioritized during the on-peak period. Even though the impact of TBP will not directly affect the detection of illegal consumers, energy consumption behavior of illegal consumers will vary significantly. Therefore, algorithms and techniques developed for detection of illegal consumers should try to quantify and consider the impact of TBP on all customers and particularly illegal consumers on the grid.
7.3 DG

DG is a small source of electric power generation or storage (typically ranging between less than a kW to tens of MW), that is not a part of a large or central power system and is located close to the load [104]. Today’s model of electricity generation and distribution in US is dominated by centralized power plants. These plants produce electricity typically by combustion (coal, water or fuel) and nuclear resources. Moreover, these power plants require a well-connected and efficient transmission system to supply electricity from generation site to customer. Centralized generation model suffers from power loss in lengthy transmission lines, issues with emission of greenhouse gasses, nuclear waste and security related issues [105].

DG can be classified into two levels: local and end-point levels. Local level DG include several site specific renewable energy sources like wind turbines, photovoltaic cells etc. The advantage with such technologies is that they are small, cost effective, reliable compared to traditional large power plants. As local level DG involves renewables, they are eco-friendly as against the centralized power plants. On the other hand, at end-point level DG every consumer can apply the above mentioned technologies. However, the most commonly used is internal combustion engine.

That being said, the driving factors for DG are classified into quality and environmental factors. DGs can improve the overall power quality as the traditional distribution system suffers with poor power quality mainly because to network switching, voltage dips, transients, interruptions, etc. Moreover, the surplus power generated by residents (using DG) can be sold to back to the grid, which can yield income during times
of peak demand. From environmentalist point of view, DG can reduce the significant amount of carbon oxides, sulfur oxides and particulate matter. Moreover, as DG technologies are independent from power grid, they can provide emergency power for hospitals, airports, military bases, prisons, police stations etc. Finally, DG can help the nation increase its diversity of energy sources [106].

7.4 Impact of DG

DG reduces the dependency of a customer on the grid and in essence, influences the pattern of energy reduced by a customer from the grid. Therefore, detection algorithms should be modified such that the reduction in customer’s dependency on the grid should not be reflected as illegal consumption. Modifications to be incorporated into the algorithms presented in chapters 4 and 6 are illustrated in Figure 7-2. In the proposed detection algorithms, it is evident that the illegal consumers are detected based on irregularities in their instantaneous energy consumption. Therefore, the impact of the DG on a customer’s consumption has to be compensated in the patterns (of irregularities) developed by encoding algorithms by carrying out the following modifications. In general, compensating the impact of DG on customer energy consumption can be done by one of the following ways:

- Collecting information about overall generation by DGs from smart meters or
- Estimate the overall consumption of each customer

However, the first option by itself may not be effective, as illegal consumers of electricity could manipulate smart meter readings related to output of DG, in addition to
manipulating the energy consumption. Therefore, to avoid this situation where DG output readings are manipulated, output from DG of a customer is advised to be estimated as proposed in Figure 7-2.

![Flowchart](image)

**Figure 7-2: Algorithm to be used for compensating the impact of DG on electricity consumption to detect illegal consumers.**

According to the algorithm, utilities should maintain reference DG sources in every geographic area such that output from that source reflects environmental and other
local conditions. Based on the output from reference sources, efficiency of other DG sources in that area can be estimated if the nameplate capacity is known, i.e., energy output from a DG source belonging to a customer at a particular point of time can be estimated. According to the algorithm, efficiency of DG should be estimated from reference DG sources. Initially, nameplate capacity of DG sources belonging to all customers on the grid has to be recorded. From the recorded values of nameplate capacity and the efficiency of DG sources calculated for that area, probable value of energy generated by DG sources belonging to a particular customer can be calculated. Independent of this, energy generated by DGs of all customers should be also collected using smart meters. Now, the collected and calculated values are compared to see if the estimated values of generations within 90-110% of the calculated value. If the values are not in the range, then average of calculated and collected values has to be finalized as the output from DG sources of a customer, if the values are in the range, then the predicted DG output value of the customer is finalized. After finalizing the value for output from DG of a customer, modifications to the magnitude of instantaneous energy consumption values received from smart meters are carried out only for customers who own DG sources. After execution of these modifications, encoding process as explained in Chapter-4 will be continued. Finally, illegal consumers are identified based on irregularities in their energy consumption.

7.5 Economic Analysis on Losses

Illegal consumption of electricity will cause financial economic loss ranging from
a few millions to billions of dollars over a year. However, it is very hard to quantity the overall financial losses due to illegal consumption of electricity owing to the following:

- Unpredictable quantity of illegal consumption.
- Unpredictable time and tariff during which illegal consumption occurs.

To analyze the impact of illegal consumption during on-peak period, consider the following load information: Date: 5/17/2012, Region: RFC, NERC. Considered “Historical Metered Load Data” is the net MWh load consumed by service regions within the PJM RTO. This data was supplied by the respective PJM Electric Distribution Companies (EDC). Load on the grid during off-peak period is around 60 MWh and reaches 80 MWh during on-peak period as illustrated in Figure 7-3. Based on this load and RTP information, money to be collected from customers can be calculated. Over that day, money to be collected ranges from $1000 per one hour during off-peak period to $4000 per hour during on-peak period as shown in Figure 7-4. Total money to be collected on that entire day is calculated to be $62,341.89.

Now, let’s consider a hypothetical scenario where illegal consumers pilfer about 15% of illegal consumption during the peak period (12:00PM to 9:00PM) and 10% of total during the off-peak period. Based on this scenario, money to be collected, and overall load on the grid, the total money lost is calculated to be $8,076.61 as displayed in Figure 7-5. The money lost is about $100/one hour during off-peak period and about $600/one hour during on-peak period.
Figure 7-3: Overall load on the grid in a service region within the PJM RTO [107].

Figure 7-4: Total money to be collected by the utility from all customers on the grid in a single day.
Figure 7-5: Money lost by the utility in a single day due to illegal consumption of electricity.

The money lost to illegal consumers is still an approximate value as the share of power generation between different types of power generation sources (Coal based generating station, nuclear power plant, renewable sources, critical generators, etc.) is unknown. Several scenarios similar to the one above may happen in the real-world. The possibility of illegal consumption at on-peak period can cause much higher financial losses than the estimates. Furthermore, impact of illegal consumption of electricity could be much serious during critical times. A critical time is a situation when the customer demand escalates and exceeds more than cumulative generation capacity of normal generators. Critical generators are operated during hot summer days, which last about 70-100 hours a year. Under such circumstances, utilities operate critical generators such as diesel based or oil fired generators. Critical generators are very expensive to operate,
about 5-15 times more expensive than conventional generators. Illegal consumption of electricity during critical times causes severe damage economically. Therefore, this analysis confirms utilities concern that actual losses will be much higher than estimates.

7.6 Summary

After implementation of smart grid, illegal consumption may become dominant during on-peak load and on-peak tariff period. With the introduction of DSM tools and DG sources make detection of illegal consumption much complicated. Eventually, perception toward illegal consumption may change overtime. To summarize Chapter 7, impact of some features and smart grid such as RTP and DG on energy consumption by illegal consumers is discussed. To this end, process involved in modifying and enhancing the existing classification algorithm is explained. In addition, details and types of RTP and DG are also illustrated.
Chapter 8

Conclusions and Future Work

This chapter concludes the work presented in the dissertation. Future work is also presented.

8.1 Conclusions

In Chapter 1, the problem statement is discussed. Dissertation objectives and organization of the dissertation are illustrated in this chapter. The importance and need for a new generalized technique for detection of illegal consumers is also discussed.

In Chapter 2, the methods implemented by illegal consumers in pilfering electricity are explained. Several issues and setbacks in implementing existing technical measures are analyzed. Several factors that influence illegal consumers to steal electricity are explained. Relevant literature review of various methods and techniques proposed and implemented for controlling illegal consumption or identifying illegal consumers.

In Chapter 3, a detailed analysis on features of smart meters and available
communication technologies are presented. Issues and challenges in the design, development, deployment, and maintenance of smart meter technologies are explained.

In Chapter 4, importance of detecting and grouping illegal consumers by analyzing customer energy consumption patterns is discussed. Procedure adopted in developing near real-time approximate energy consumption data for a wide range of customers is illustrated. A novel data encoding technique is proposed to simplify the analysis of energy consumption data by mapping instantaneous energy consumption readings from smart meters into irregularities in energy consumption by each customer. Detailed procedure involved in implementing the encoding technique is presented. In addition to the faster analysis, the encoding technique enhances the entire process of detecting illegal consumers.

In Chapter 5, advanced data classification algorithms for classifying illegal consumers based on their energy consumption patterns are elucidated. These methods are designed to detect illegal consumers in a smart grid environment where utilities possess access to customers’ real time energy consumption data. SVM model, GA based NN model, Rule Engine algorithms, and NNPR tools have been developed to classify and group illegal consumers. The process, order and priority in which these tools are modeled and used are explained in this chapter.

In Chapter 6, the importance of implementation of algorithms for detection of illegal consumers on High Performance Computing is discussed. Data
loading, encoding presented in chapter 4 and classification algorithms presented in chapter 5 are modified to explore the advantages of parallel programming and processing. The newly developed algorithms supporting Task Parallelism and Data Parallelism approaches are also illustrated.

In Chapter 7, impact of some features and smart grid such as RTP and DG on energy consumption by illegal consumers is discussed. To this end, process involved in modifying and enhancing the existing classification algorithm is explained. In addition, details and types of RTP and DG are also illustrated.

8.2 Possible Future Research Directions

Reducing the classification time:

A specific set of customers (probable illegal consumers) are selected using Artificial Intelligence (AI) tools or optimization techniques, or a set of specific inputs (instances of time) are selected using feature selection algorithms for faster identification of illegal consumers. As a first step in every attempt of classification, energy consumption profiles of selected customers or all customers just at selected time-instances may be tested first.

Encoding technique can be extended such that the total number of irregularities (32 inputs) per customer for a single day can be reduced to 19 inputs for the same day. This can be done by grouping 5 energy readings instead of 3 during the encoding process. On the flip side, the maximum value of an encoded input rises to 31. For the classification with Rule Engine
algorithm, this maximum number will not have much impact on the performance of the classification model because of the fact that the rules used in the Rule Engine can be easily altered to meet the requirement. Therefore, the number of inputs to be combined (3, 4 or 5) can be decided depending on the manageable capability and complexity of the classification algorithm.

Enhancing the performance and scope of the problem:

- Encoding algorithm can be enhanced by introducing more real-world parameters and variables (both technical and non-technical), such that the process of mapping energy patterns into irregularities is further strengthened. Integration of these new technical and non-technical parameters into classification algorithms may come into light after complete implementation of smart grid. This inclusion will ensure the impact of such parameters on algorithms for identification of illegal consumers.

- Enhancement of the analysis on the impact of TBP and DG on customer energy consumption patterns in collaboration with DSM and smart home management tools can be done. Study on the impact of TBP and RTP on illegal consumption of electricity will be one of the most significant factors of future home energy management. This analysis will benefit both genuine customers and utilities.

- Including the impact of illegal consumption in tools that perform dynamic optimization of Voltage and Vars are used to reduce the load on the grid.

- Integration of Geographic Information System (GIS) for real-time visualization of a customer’s energy consumption, pattern of irregularities,
and power generated from DG sources on customer premises. This portion of the work can be further extended to thematic mapping, e.g. economy and illegal consumption of electricity or literacy and illegal consumption of electricity. Thematic mapping provides an extensive understanding of the impact of illegal consumption on a large geographic entity.

On the other hand, illegal consumption of electricity may be controlled by focusing on cyber security (designing stronger firewalls or enhancing firmware in view of cyber security). For example, complicating the process of meter tampering and reducing the hackings of smart meters through the establishment of certain standards. These standards are envisioned to control the flexibility of illegal consumers in installing external devices or software or firmware updates from third-party developers on their smart meters.
References


Appendix-A

List of Papers Submitted Based on this Dissertation


