Enhancing Health and Energy Efficiency Through Data-Driven Urban Initiatives: A Smart City Approach

by

Jin Soo Cho

Mina Sartipi
Professor of Computer Science
(Chair)

Nancy Fell
Professor of Physical Therapy
(Committee Member)

Dalei Wu
Professor of Computer Science
(Committee Member)

Lani Gao
Professor of Mathematics
(Committee Member)

Enhancing Health and Energy Efficiency Through Data-Driven Urban Initiatives: A Smart City Approach

by

Jin Soo Cho

A Thesis Submitted to the Faculty of the University of Tennessee at Chattanooga in Partial Fulfillment of the Requirements of the Degree of Ph.D. in Computational Science: Computer Science

The University of Tennessee at Chattanooga
Chattanooga, Tennessee

May 2021

ABSTRACT

With the recent growing recognition, the "smart city" project aims to advance the quality of modern cities through technology and data science. In this dissertation, two fundamental smart city applications are explored: Smart Health and Smart Energy. The goal of the presented studies is to transform the future of healthcare and energy through data-driven solutions. For Smart Health, statistical analysis and machine learning algorithms are employed to improve patient management and their eventual outcomes. This is done by implementing a predictive analytics framework to identify various risk factors associated with respective medical conditions. The aim of the Smart Energy application is to analyze energy meter data to improve energy efficiency and manage power demand in both residential and industrial sectors. Various state-of-the-art machine learning algorithms are investigated by scrutinizing data obtained from multiple sources. The proposed method introduced in this dissertation emphasizes the effectiveness of data-driven approaches in urban development and planning. The unification of technology and infrastructure will improve individual quality of life and advance the community into a new era of smart society.

ACKNOWLEDGMENTS

I would like to thank many people who have guided me to get to this point. First and foremost, I want to express my gratitude to my parents, Chong and Ok Cho, as well as my brothers, Jin Young Cho and Sam Nho. They gave me so much encouragement and energy for me to finish this long journey. This would have not been possible without the support from my family.

I am extremely grateful to my supervisor, Dr. Mina Sartipi. Her experience and knowledge have encouraged me in all the time of my academic research and throughout my academic career. She has been my mentor and supervisor for the past 6 years. During these years, she has provided me with valuable research experience by demanding a high quality of work, supporting my attendance at various conferences, and providing constructive feedback. My academic and research career would have not been possible if it wasn't for her help.

I would like to thank the previous members of SCAL. The previous members include: Dr. Zhen Hu, Brandon Allen, Brian Williams, and Austin Harris. Dr. Zhen Hu has been a huge part of my research life. Also, I would like to thank all the current members of CUIP.

I would also like to thank the faculty in the Computer Science Department at the University of Tennessee at Chattanooga. They have taught me valuable lessons and pushed me to succeed and focus on my future.

Finally, I would like to acknowledge my dissertation committee members: Dr. Nancy Fell, Dr. Dalei Wu, and Dr. Lani Gao for their support and guidance throughout my academic journey. Thank you all for your support and everything that you have done for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Background

The idea of the smart city covers a broad range of disciplines that leads to urban life that is safe, environmentally secure, and efficient in every aspect whether for energy, transportation, healthcare, and etc [1]. This is done through the utilization of advanced computing resources, integrated materials, sensors, databases, and the state-of-the-art algorithms [1,2]. The city of Chattanooga has been working closely with the Centers for Urban Informatics and Progress (CUIP) to transform the city into an ubiquitous city. With the recent work in many domains, Chattanooga is considered one of the cities that bear core factors for a successful smart city initiatives [3]. Sustainable and innovative smart city can be achieved through a health-centric data-driven urban initiatives that integrates health, energy, transportation, computer science, and data science. Two key smart city initiatives are investigated in this dissertation: Smart Health and Smart Energy.

1.2 Smart Health

Expeditious advancement of Internet of Things (IoT), information communication technologies (ICT), machine learning techniques, and smart technology solutions have introduced the ability to revolutionize from reactive healthcare to preventive, proactive, and decision-based healthcare. Healthcare-related innovations such as remote patient monitoring, telemedicine, data-based public health interventions, and integrated patient manage-

ment systems can substantially impact quality of life and long-term health outcomes. In addition to our previous work on mStroke, which uses wireless sensor technology to establish a remote extended monitoring and mobile health system for risk-related stroke measures, electronic medical records (EMR) data is analyzed to support data-driven smart health approach and enable post-stroke health management through data analytics. EMR data will offer an opportunity to access the historical and recent data that will have a unique clinical care asset for the future diagnosis and treatments, opening opportunities to personalized medicine, preventive care, and chronic disease management. For the post-stroke management, hospital discharge disposition is investigated in order to support individual's aging in place. Moreover, readmission rate is explored to identify individuals post-stroke with the highest risk of stroke recurrence and/or re-hospitalization, which can help to triage work lists and focus healthcare efforts on patients that are truly high-risk.

In addition to the post-stroke management, Natural Language Processing (NLP) technique is explored to assist in clinical matching recommendations in order to properly recommend medical trials for specific cancer patients based on their conditions.

1.3 Smart Energy

Energy is one of the most important factors in solving the modern issues such as climate change, health, global energy, environment, and sustainable development [4]. Smart energy was introduced in effort to transform the traditional energy system to the future sustainable energy system. The goal of the smart energy initiative is to demonstrate and achieve the most affordable and efficient ways to implement future sustainable energy framework within the context of the smart city [5]. The last decades have observed severe changes in weather patterns and the societal damages. Significant growth of power consumption, mostly for cooling purposes, and possible failure to meet the energy needs due to fuel shortage and capacity limitations have shown to lead to failure in operation and availability of critical

infrastructures. Hence, smart energy initiatives could ensure power system resiliency, reliability, and availability during the course of extreme temperature events, and guarantee the supply of power to critical loads. In addition, optimizing the performance of assets can prolong their lifetime and postpone the need for overhaul and replacements. This promotes power grid sustainability and can indirectly lower the operational costs of the network which may benefit both the electric utilities and the end users. Thus, identifying the activity level of the A/C unit at individual residential units is crucial in properly managing demand response at the building. Moreover, energy anomaly patterns are investigated in order to provide proper management of meter process in the industrial settings. This will allow for correct identification of load transfers, meter damage, and outages to minimize error and ensure accurate energy billing.

1.4 Research Objective

The main objective of this dissertation is to incorporate data science and machine learning to investigate different smart city problems. Smart Health projects aims to address medical research problems such as predicting hospital discharge disposition status, analyzing readmission rates, predicting patient outcomes using machine learning algorithms, and implementing clinical trial matching recommendation system. On the other hand, Smart Energy projects aims to investigate efficient energy usage in the residential and industrial sectors. Through these projects, energy usage patterns of both residential and industrial sectors are studied and used as a blueprint for the enhancement and development of future smart cities.

1.5 Thesis Layout

The remainder of this work is structured as followed: Chapter 2 discusses the development and validation of a simple predictive tool for determining stroke patients hospital discharge disposition status based on statistical analysis. Chapter 3 extends the validation process of the easy-to-use risk score calculator by analyzing 30-day readmission rates based on the data provided by the Centers for Medicare and Medicaid Services (CMS). Chapter 4 discusses the strengths of different black-box models in predicting medical outcomes and the interpretation of the results using Local Interpretable Model-agnostic Explanations (LIME). Chapter 5 introduces the use of machine learning algorithms and natural language processing (NLP) techniques to develop a clinical trial matching system for cancer patients. In this chapter, two major components of the proposed system are discussed: conditional logic process and clustering. Chapter 6 discusses the use of time-series analysis techniques and deep learning methods to forecast A/C electric energy usage based on residential smart meter data. In this chapter, several energy models are developed to fully understand how energy is used in different houses, and the impact of controlling the A/C unit on the temperature inside the buildings is investigated. Chapter 7 investigates two major algorithms: LSTM & ARIMA for time series analysis. The combination of these two models are used to forecast energy demand with the intent of benefiting from the strengths from the two models. Through this work, energy anomaly detection system is implemented to create a automated tool that detects unusual energy behavior in the industry setting. Lastly, Chapter 8 provides a quick summary of the smart city projects and their role in the smart city.

CHAPTER 2

STATISTICAL ANALYSIS BASED SMART HEALTH

2.1 Introduction

Stroke is the fifth leading cause of death and a leading cause of long-term disability in the United States, where each year approximately 800,000 people experience a stroke, including 610,000 new and 185,000 recurrent strokes, at a cost of $34 billion [6, 7]. The state of Tennessee lies in the "Stroke Belt" of the United States and has the highest prevalence of stroke and its corresponding risk factors [8, 9]. Patients with significant physical, cognitive, and/or behavioral deficits after stroke often are referred for intensive rehabilitation. Early research suggests that the site for postacute stroke care (eg, inpatient rehabilitation facility [IRF], skilled nursing facility (SNF), home with/without home health (HH), or outpatient rehabilitation services) significantly affects 6-month functional outcomes in the domains of basic mobility, activities of daily living, and applied cognition. In a study in northern California, patients who went to an IRF postacute stroke had better functional outcomes than those who received care through an SNF, HH, or outpatient rehabilitation services [10]. Yet clinicians and discharge planners continue to grapple with the lack of standardized assessment capable of predicting optimal postacute discharge resource allocation. Furthermore, the rehabilitation needs assessment and the subsequent insurance approval process can take days, thereby resulting in an unnecessary longer hospital stays and potentially exposing patients to hospital-acquired infections. Early determination of hospital discharge disposition, especially at an acute admission, if possible, can optimize acute stroke care at the hospital, help with prognostication, allow sufficient time for patients and their families to prepare for

postacute stroke care, and provide sufficient time for finding the appropriate rehabilitation program and obtaining the requisite insurance approval [11]. As such, early identification of discharge disposition may be extremely important for stroke management, decision support, and eventual outcomes for patients with stroke.

2.2 Related Work

Although several studies have examined patient characteristics associated with hospital discharge disposition, the results of these studies are inconsistent [12–23]. A literature review of 19 articles found that functional dependence, comorbidity, neurocognitive dysfunction, previous living circumstances, and marital status were significantly associated with other than home discharge for patients with stroke [23]. The effect of age, sex, race, affected hemisphere, or availability of a caregiver on hospital discharge disposition was inconsistent across studies, however [23]. Furthermore, few studies have proposed a discharge disposition predictive model for use in acute patients with stroke. A discharge disposition predictive model after acute stroke using the Taiwan Stroke Registry data with 21,575 patients with stroke was reported but lacked generalizability to populations outside Taiwan and used clinical parameters that may not be available at the time of a patient's presentation with stroke [19]. In the United States, the Northeast Cerebrovascular Consortium piloted a formal rehabilitation needs assessment with discharge referral prediction in the acute hospital setting. They determined that the sociodemographic characteristics, premorbid function, and Barthel Index activities of daily living score for patients with stroke discriminated between discharge home and inpatient rehabilitation (SNF and IRF) [24].

As such, our goal was to develop and validate a simple predictive tool for determining hospital discharge disposition status using easily available patient characteristics (sex, age, race, stroke type, comorbidity, source of admission, primary payer class, and secondary payer class) at the time of a patient's presentation with acute stroke symptoms. To meet our goal,

we evaluated the association of patient characteristics with hospital discharge disposition status based on the data provided by the Tennessee Department of Health through the Hospital Discharge Data System.

2.3 Methods

2.3.1 Study Population

We used data from the Hospital Discharge Data System maintained by the Tennessee Department of Health. The purpose of the Hospital Discharge Data System is to collect and summarize hospital claims data and to analyze and compare charges for similar types of services [25]. The dataset included all of the records of hospitalized patients with the principal diagnosis of stroke (International Classification of Diseases, Ninth Revision codes 430, 431, 433, 434, and 436). The dataset contains information on patient demographics, primary and secondary diagnoses, procedures performed, and insurance status.

2.3.2 Variables

We stratified age into three categories: 18 to 64 years, 65 to 74 years, and 75 years and older and stroke types were pooled into three categories: ischemic, subarachnoid hemorrhage, and intracerebral hemorrhage. We included diabetes mellitus, heart disease, hypertension, peripheral arterial disease, chronic kidney disease, hyperlipidemia, arrhythmia, and depression as comorbid conditions. Sources of patient referrals to hospital were grouped into home or a nonhealthcare facility, clinic or physician's office, or another hospital. Health insurance was categorized into private insurance, Medicaid, Medicare managed, and Medicare fee-for-service.

Discharge disposition status was defined as home discharge when patients were discharged home with or without HH care services and as facility discharge when patients were discharged to healthcare facilities such as an SNF, an intermediate care facility, IRF, and another short-term general hospital for inpatient care [25].

2.3.3 Statistical Analysis

Demographic and clinical characteristics of patients with stroke with home discharge were first compared with facility discharge counterparts using Pearson Chi Square test. To develop our predictive tool, we divided the whole dataset into a derivation cohort and a validation cohort. The derivation cohort consisted of records of patients with stroke hospitalized from 2010 through 2013 and the validation cohort consisted of records of patients with stroke hospitalized in 2014. Based on the derivation cohort, logistic regression was performed to estimate odds ratios (ORs) of patient characteristics associated with facility discharge. Both unadjusted and adjusted ORs with 95% confidence intervals (CIs) were considered. Next, coefficients from the multivariable logistic regression related to adjusted ORs were used to derive risk scores [26, 27]. A total risk score was calculated for each patient by adding corresponding risk scores [26, 27]. Following the logistic function, the predicted probability of facility discharge for each total risk score was given and compared with the observed counterpart. Eventually, an easy-to-use predictive tool was built by using the total risk score to predict the hospital discharge disposition status of each patients with stroke. We assessed the performance of such a predictive tool using the receiver operating characteristic (ROC) curve and the area under a ROC curve (AUC) with 95% CI.

2.3.4 Results

The original dataset for our investigation included 139,706 records of patients with the principal diagnosis of stroke, hospitalized from 2010 to 2014. We excluded 12,125 records (invalid or missing data: 1151, deceased/expired: 6855, discharged to hospice: 3185, discontinued care/court: 934). Of the remaining 127,581 records, 86,114 (67.5%) were related to home discharge and 41,467 (32.5%) corresponded to facility discharge (Table 2.1). All of the examined patient characteristics were significantly associated with hospital discharge disposition status (Table 2.1). The ratios of patients with stroke discharged to a facility compared with home remained stable during the study period (2010: 0.51, 2011: 0.54, 2012: 0.52, 2013: 0.54, 2014: 0.55).

Table 2.1 Demographic and clinical characteristics of stroke patients

Characteristics	Home Discharge (n = 86,114)	Facility Discharge (n = 41,467)	P value
Sex			<0.0001
Men	43,955 (51.0%)	18,708 (45.1%)	
Women	42,159 (49.0%)	22,759 (54.9%)	
Age			<0.0001
18-64	36,136 (41.9%)	13,604 (32.8%)	
65-74	25,673 (29.8%)	9,896 (23.9%)	
≥75	24,305 (28.3%)	17,967 (43.3%)	
Race			<0.0001
White	71,469 (82.9%)	33,114 (79.9%)	
Black	11,533 (13.4%)	7,012 (16.9%)	
Other	3,112 (3.7%)	1,341 (3.2%)	
Stroke Type			<0.0001
Ischemic	78,774 (91.5%)	34,143 (82.3%)	
Subarachnoid hemorrhage	3,184 (3.7%)	2,383 (5.8%)	
Intracerebral hemorrhage	4,156 (4.8%)	4,941 (11.9%)	
Comorbidity			
Diabetes	21,353 (24.8%)	14,357 (34.6%)	<0.0001
Heart disease	30,237 (35.1%)	21,205 (51.1%)	<0.0001
Hypertension	48,877 (56.8%)	32,055 (77.3%)	<0.0001
Peripheral arterial disease	5,831 (6.8%)	2,120 (5.1%)	<0.0001
Chronic kidney disease	6,004 (7.0%)	5,322 (12.8%)	<0.0001
Hyperlipidemia	27,892 (32.4%)	15,006 (36.2%)	<0.0001
Arrhythmia	10,150 (11.8%)	10,766 (25.9%)	<0.0001
Depression	4,730 (5.5%)	3,486 (8.4%)	<0.0001
Source of Admission			<0.0001
Non-healthcare facility	56,752 (65.9%)	30,788 (74.2%)	
Clinic or physician's office	19,134 (22.2%)	1,696 (4.1%)	
Transfer from a hospital	6,014 (6.9%)	4,544 (10.9%)	
Others	4,214 (5.0%)	4,439 (10.8%)	
Primary Payer Class			<0.0001
Medicare (Not managed)	40,441 (46.9%)	23,645 (57.0%)	
Medicare (Managed)	14,172 (16.5%)	6,740 (16.3%)	
Medicaid	633 (0.7%)	262 (0.6%)	
Private Insurance	23,021 (26.7%)	7,586 (18.3%)	
Others	7,847 (9.2%)	3,234 (7.8%)	
Secondary Payer Class			<0.0001
Medicare (Not managed)	6,327 (7.3%)	3,042 (7.3%)	
Medicare (Managed)	2,143 (2.5%)	1,162 (2.8%)	
Medicaid	5,725 (6.6%)	4,302 (10.4%)	
Private Insurance	24,133 (28.0%)	12,379 (29.9%)	
Others	47,786 (55.6%)	20,582 (49.6%)	

The derivation and validation cohorts included 101,223 and 26,358 records, respectively (size ratio: 3.8:1). Based on both unadjusted and adjusted ORs, patient characteristics such as female sex; ages 75 years and older; black race; a subarachnoid or intracerebral hemorrhage; presence of diabetes mellitus, hypertension, heart disease, chronic kidney disease, arrhythmia, or depression; fee-for-service Medicare; and transfer from an outside hospital were associated with an increased risk of having a facility discharge (Table 2.2).

Table 2.2 Odds ratios of patient characteristics associated with facility discharge

Characteristics	Unadjusted Odds Ratios	Adjusted Odds Ratios	β
Sex			
Men	1.00 (Ref.)	1.00 (Ref.)	0
Women	1.27 (1.24-1.30)	1.15 (1.12-1.19)	0.1427
Age			
18-64	1.00 (Ref.)	1.00 (Ref.)	0
65-74	1.02 (0.98-1.05)*	1.01 (0.96-1.06)*	0.0129
$\geq$75	2.00 (1.94-2.06)	1.81 (1.72-1.91)	0.5955
Race			
White	1.00 (Ref.)	1.00 (Ref.)	0
Black	1.31 (1.26-1.36)	1.15 (1.11-1.20)	0.1440
Other	0.92 (0.86-0.99)*	0.78 (0.73-0.85)	-0.2428
Stroke Type			
Ischemic	1.00 (Ref.)	1.00 (Ref.)	0
Subarachnoid hemorrhage	1.72 (1.62-1.83)	2.34 (2.19-2.50)	0.850
Intracerebral hemorrhage	2.78 (2.65-2.92)	2.91 (2.76-3.07)	1.068
Comorbidity			
Diabetes	1.22 (1.19-1.26)	1.29 (1.25-1.34)	0.2563
Heart disease	1.14 (1.09-1.18)	1.15 (1.11-1.20)	0.1433
Hypertension	2.31 (2.23-2.39)	1.90 (1.84-1.97)	0.6438
Peripheral arterial disease	0.59 (0.55-0.62)	0.68 (0.63-0.72)	-0.3922
Chronic kidney disease	1.37 (1.31-1.44)	1.24 (1.18-1.30)	0.2127
Hyperlipidemia	0.77 (0.75-0.79)	0.83 (0.81-0.86)	-0.1815
Arrhythmia	2.08 (2.00-2.17)	1.71 (1.64-1.79)	0.5373
Depression	1.38 (1.30-1.45)	1.35 (1.28-1.43)	0.3024
Source of Admission			
Non-healthcare facility	1.00 (Ref.)	1.00 (Ref.)	0
Clinic or physician's office	0.16 (0.15-0.17)	0.20 (0.19-0.21)	-1.6026
Transfer from a hospital	1.46 (1.39-1.53)	1.18 (1.13-1.24)	0.1684
Others	1.98 (1.89-2.07)	1.73 (1.65-1.82)	0.5485
Primary Payer Class			
Medicare (Not managed)	1.00 (Ref.)	1.00 (Ref.)	0
Medicare (Managed)	0.80 (0.77-0.83)	0.75 (0.72-0.79)	-0.2883
Medicaid	0.75 (0.64-0.88)	0.71 (0.60-0.86)	-0.3362
Private Insurance	0.56 (0.54-0.58)	0.72 (0.68-0.75)	-0.3351
Others	0.71 (0.67-0.74)	0.72 (0.67-0.77)	-0.3275
Secondary Payer Class			
Medicare (Not managed)	1.00 (Ref.)	1.00 (Ref.)	0
Medicare (Managed)	1.17 (1.07-1.27)	1.14 (1.04-1.26)	0.1331
Medicaid	1.60 (1.50-1.71)	1.67 (1.55-1.80)	0.5148
Private Insurance	1.10 (1.04-1.16)	1.10 (1.04-1.17)	0.0945
Others	0.92 (0.88-0.97)	1.34 (1.26-1.42)	0.2928

$*$: $p \geq 0.05$

The range of the calculated risk scores for patient characteristics was from -14 to 9 (Table 2.3). The range of the total risk score for a given patient was from -20 to 39. The predicted probability of facility discharge increased with the total risk score following logistic function (Fig 2.1), which means a patient with a higher total risk score had a higher chance of being discharged to a healthcare facility. Because the number of patients with total risk scores of >30 was small, only results corresponding to total risk scores from -20 to 30 were reported. Furthermore, the observed probabilities of facility discharge for both derivation and validation cohorts were consistent with the predicted counterpart (Fig 2.1).

To confirm the usefulness of the easy-to-use predictive tool, ROC curves for both derivation and validation cohorts were plotted (Fig 2.2). The corresponding AUCs of the derivation and validation cohorts were 0.737 (95% CI 0.734-0.740) and 0.724 (95% CI 0.718-0.730), respectively. We simplified risk scores further and considered only five patient characteristics (sex, age, race, stroke type, and comorbidity) and exploited only two positive integers (1 or 2) to represent risks (Table 2.3). When such simplified risk scores were applied, the AUCs of the derivation and validation cohorts were 0.693 (95% CI 0.689-0.696) and 0.679 (95% CI 0.673-0.686), respectively (Fig 2.2).

Table 2.3 Risk scores of patient characteristics associated with facility discharge

Characteristics	Risk Score	Simplified Risk Score
Sex		
Men	0	0
Women	1	1
Age		
18-64	0	0
65-74	0	0
≥75	5	2
Race		
White	0	0
Black	1	1
Other	-2	0
Stroke Type		
Ischemic	0	0
Subarachnoid hemorrhage	7	2
Intracerebral hemorrhage	9	2
Comorbidity		
Diabetes	2	1
Heart disease	1	1
Hypertension	6	2
Peripheral arterial disease	-3	0
Chronic kidney disease	2	1
Hyperlipidemia	-2	0
Arrhythmia	5	2
Depression	3	1
Source of Admission		
Non-healthcare facility	0	-
Clinic or physician's office	-14	-
Transfer from a hospital	1	-
Others	5	-
Primary Payer Class		
Medicare (Not managed)	0	-
Medicare (Managed)	-3	-
Medicaid	-3	-
Private Insurance	-3	-
Others	-3	-
Secondary Payer Class		
Medicare (Not managed)	0	-
Medicare (Managed)	1	-
Medicaid	4	-
Private Insurance	1	-
Others	3	-

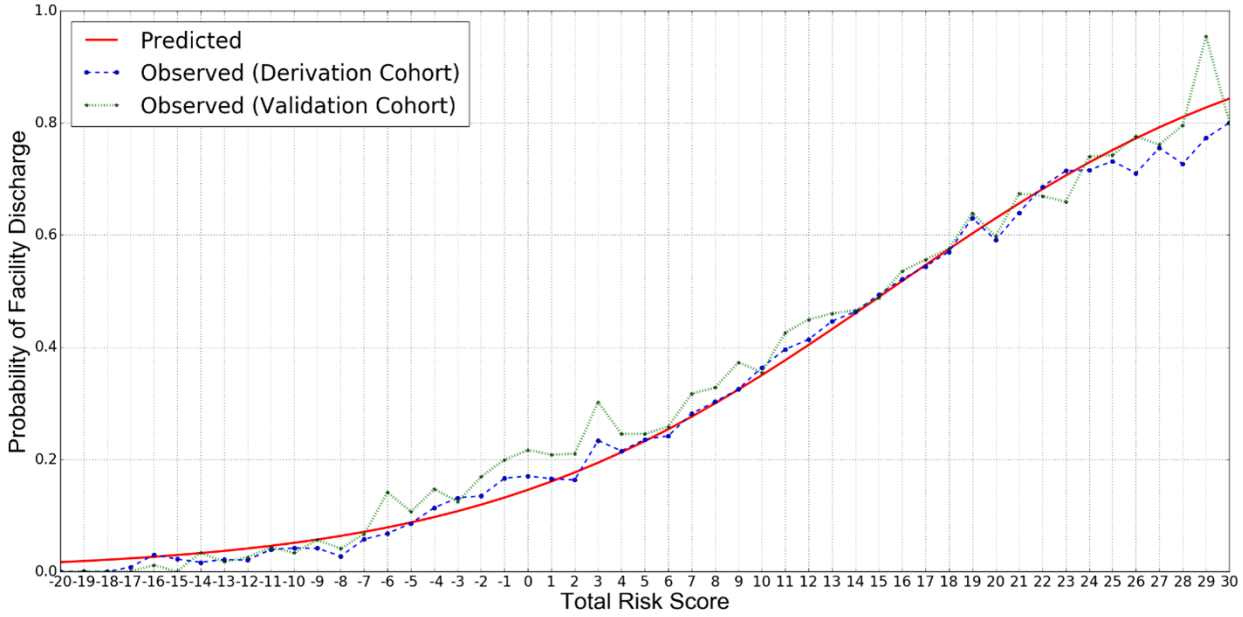


Figure 2.1 Predicted and observed probabilities of facility discharge for each total risk score

2.4 Discussion

In this study we developed and validated a discharge disposition predictive tool based on integer-based risk scores for patients hospitalized with a principal diagnosis of stroke. This easy-to-use tool had a significant discriminatory capability and used patient characteristics available at the time of a patient's presentation to a hospital. The hospital discharge disposition results from multiple factors with mixed effects, so risk scores were derived from coefficients of multivariable logistic regression related to an adjusted OR. Based on the adjusted OR, the top five patient characteristics associated with a high risk of facility discharge were identified as an intracerebral hemorrhage, a subarachnoid hemorrhage, hypertension, ages 75 years and older, and arrhythmia.

We identified a strong correlation between hospital discharge disposition and the studied patient characteristics, which aligns with the findings of other investigators. We found that female patients with stroke in Tennessee were more likely than others to be discharged to

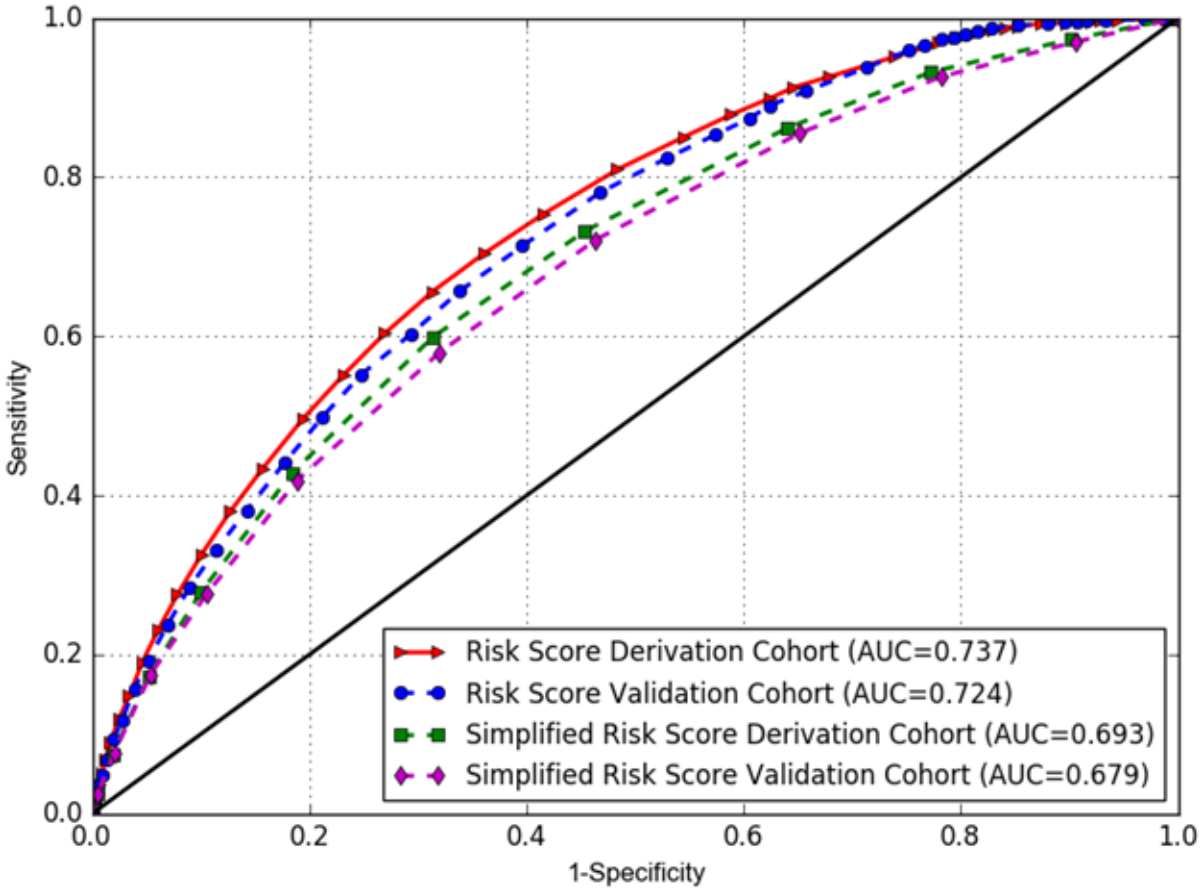


Figure 2.2 ROC curves of risk scores and simplified risk scores

a facility rather than home [14]. A study mentioned that patients' marital status and sex could play a role in institutionalization [28]. One of the reasons that female patients are more likely to be discharged to a facility is that male caregivers are less experienced in providing care to their spouses in comparison with female caregivers. This also aligns with the fact that patients receiving inadequate support from their caregivers often are discharged to a facility or other institutions [29].

Older patients also were more likely be discharged to a facility rather than home [14,19,30]. In our study the probability of patients with stroke being discharged to a facility increased as their age increased. For example, for the age group of 75 years and older (OR 1.81, 95% CI 1.72-1.91), the probability of being discharged to a facility nearly doubled compared with

the age group of younger than 64 years (OR 1.00, Reference). Because it is hard for them to take care of their own health [31], they must rely on systematic and careful management from the facility. Furthermore, as patients with stroke age, their caregivers may have a different condition and may be unable to provide adequate support for them [19].

Patients with hemorrhagic stroke also were likely to be discharged to a facility rather than home. Compared with ischemic stroke, hemorrhagic stroke is much more severe because of a higher mortality rate and different medical procedures [32]. As such, patients with hemorrhagic stroke need attentive care from institutions.

We also have found that African American/black patients were more likely than patients of other races to be discharged to facilities [14]. This finding may be confounded by a lower socioeconomic status such as education, working status, and household income [33].

Patients with stroke and diabetes mellitus [19], heart disease [34], hypertension, chronic kidney disease [35], arrhythmia [34], or depression were more likely than others to receive care from a variety of healthcare facilities. Based on the Tennessee hospital discharge data, the prevalence of hypertension among patients with stroke was high and consequentially led to an increased odds for discharge to another facility. This finding reflects the overall higher prevalence of hypertension among Tennessean adults compared with national estimates [36, 37] and highlights the need for early detection and control of this important risk factor within the state's adult population [6, 9, 38]. The OR of facility discharge for individuals with poststroke arrhythmia was high, even though the population was small (10,150; OR 1.71, 95% CI 1.64-1.79) compared with other major comorbidities such as diabetes mellitus (21,353; OR 1.29, 95% CI 1.25-1.34), heart diseases (30,237; OR 1.15, 95% CI 1.11-1.20), or hyperlipidemia (27,892; OR 0.83, 95% CI 0.81-0.86). This result suggests the need for further investigation on the correlation between arrhythmia and stroke.

In this article we presented a tool developed with a focus on clinical utility and the rapidity of discharge disposition determination. The predictive tool has important clinical implications because it may serve as a strong first assessment for acute stroke discharge

disposition analysis in the acute hospital setting. As discussed in the comparison to two other models reported in the literature [19, 39], our tool is simple, can be implemented by a healthcare provider with minimal training, and can provide guidance to care coordinators at the time of admission in preparing for an adequate discharge disposition. Early discharge planning is not only associated with decreased duration of acute hospitalization but also with improved patient-centered outcomes such as decreased readmission rate and duration [40]. Furthermore, early patient transition to optimal discharge disposition reduces costs [41]. European hospitals are studying early supported poststroke discharge and exploring key patient variables such as premorbid functional status and cognitive function [42]. Likewise, our predictive tool can be used for future research to identify patient subsets that can benefit from early discharge to home [42].

Our investigation is subject to at least two limitations. First, we did not have information about the functional status of the patients with stroke, subsequent to the stroke. The functional/behavioral measures such as the National Institutes of Health Stroke Scale (NIHSS), the Functional Independence Measure, the Barthel Index, and the Rankin Scale were not available through these hospital discharge data. Having access to any of these measures would have strengthened the final models in our analysis and would have further aided our assessment of stroke severity and its correlation with discharge disposition status [20, 23, 43]. For example, others have shown that the NIHSS score at admission is a potential factor for discharge disposition prediction, in which the corresponding AUC can be as high as 0.84 [19]. The NIHSS score also has been used for risk adjustment to determine racial and ethnic differences in clinical outcomes [17]. Our findings of selected sociodemographic, clinical, and insurance status being strongly associated with the prediction of hospital discharge disposition align well with the studies in which functional/behavioral measures are included, however. A second limitation is that the Tennessee hospital discharge data did not allow us to differentiate stroke care among patients by primary hospital. This may be an important confounder for our findings because other investigators have shown significant variability in

stroke outcomes by hospital facility, where teaching hospitals and certified stroke centers reported better stroke outcomes compared with community hospitals [44].

In sum, our investigation of hospital discharge disposition in the State of Tennessee suggests significant benefit and effectiveness in promoting both pre-clinical research and clinical utilization of stroke management and decision-making support. Proactive intervention, targeted treatment, and personalized care planning for stroke patients can be enabled with the early determination of hospital discharge disposition at an acute stroke admission. Furthermore, our predictive tool, which is based on simple risk scores, may be an attractive and easily adoptable discharge risk tool for use by physicians or nurses in clinical practice, which may assist with an early discharge disposition prediction and become a standard procedure or health service in stroke management and decision support. Further study is required to determine whether our discharge disposition predictive tool may be expanded to discriminatively predict between SNF and IRF placement options and long-term patient outcomes.

2.5 Conclusion

The early determination of hospital discharge disposition status at an acute stroke admission is highly valuable for stroke management and can optimize stroke system of care. Our study discovered the hospital discharge disposition pattern of stroke patients in Tennessee and identified top five patient characteristics associated with a high risk of facility discharge as an intracerebral hemorrhage, a subarachnoid hemorrhage, hypertension, age $\geq$ 75 years, and arrhythmia. Based on our findings, we have developed an easy-to-use predictive tool using the derived integer-based risk scores. This tool can be adopted for such an early and quick determination by physicians or nurses in clinical practice.

CHAPTER 3
30-DAY READMISSION VALIDATION

3.1 Introduction

Determining discharge disposition after stroke is a complex decision-making process by the healthcare team. After index hospitalization in a short-term acute care hospital, patients may be discharged to their home or another facility for continued medical or rehabilitative management. Many factors affect a patient's discharge destination, including patient-related factors such as age, race, comorbidities, and functional status [23] as well as healthcare system-related factors such as bed availability and workforce [11]. In the acute care hospital, the healthcare team works together with the patient and the patient's family to determine whether the patient can return home or requires transfer to another facility. The site of post-acute care has effects on overall mortality [45] and 6-month functional outcomes in the domains of basic mobility, activities of daily living, and applied cognition [10]. The process of determining discharge destination is often delayed by insurance approval, rehabilitation assessment, and medical management, thus increasing the patient's length of stay, risk of infection, and unnecessary costs. Early prediction of discharge destination may optimize post-stroke care and improve outcomes by mitigating these delays. While many have attempted to predict discharge disposition after stroke [19, 23, 39, 46], outcomes are limited to validate whether the prediction was truly appropriate for the patient in a clinically meaningful way.

Hospital readmission is one metric of quality of care and discharge planning. Low readmission rates indicate the proper and thorough care with appropriate discharge disposition.

Readmissions are costly to the healthcare system, averaging $14,400 per readmission and affecting 13.9% of all index hospitalizations [47]. In Medicare beneficiaries, 30-day readmission rates approached 20% at an estimated cost of $17.4 billion in 2004 [48]. With the Centers for Medicare and Medicaid Services (CMS) Hospital Readmissions Reduction Program (HRRP), hospitals receive reduced payment for services rendered for excess readmissions [49]. As this program expands [50], hospitals continue to strive to identify and address preventable readmissions. Predictive analytics is one strategy being used with other conditions to mitigate excess readmissions and reduce cost by identifying and intervening for patients who are at a high risk of readmission [51–53]. After a stroke, patients are at high risk for complications such as recurrent stroke, fractures, deep vein thrombosis, and urinary tract infections [54]. 30-day readmission rates for stroke range from 8.7% to 17.4% [55–57]. Preventing these readmissions is one of the primary goals of discharge planning in the acute care hospital. In this study, readmission status was used as a measure of the clinical significance and effectiveness of a discharge disposition prediction tool.

The purpose of this research is to create and validate a predictive tool for discharge disposition post-stroke in Medicare beneficiaries from 2014 and 2015 claims. Most strokes occur in people over age 65 [58]; therefore, CMS data is well-suited for studying this patient population. The predictive tool aimed to develop a risk score for each patient based on demographics related to stroke risk and clinical characteristics at the point of the index hospitalization. We hypothesized that patients with a higher risk score would have a higher chance of being discharged to a healthcare facility. Validation of the predictive tool was based on readmission rates when the prediction differed from the patient's actual discharge location. We hypothesized that there would be higher readmission rates when a patient was discharged home but the prediction tool recommended discharge to a facility for continued management.

3.2 Methods

3.2.1 Data

We used data from the Centers for Medicare and Medicaid Services (CMS). Our dataset includes all of the records of hospitalized patients with the principal diagnosis of stroke (*International Classification of Diseases, Ninth Revision* codes 430, 431, 432,0, 432.1, 432.9, 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, 434.91, 435, 435.0, 435.1, 435.3, 436, 437.1, 437.5, 997.02). This data contains information such as patient demographics, diagnosis codes, procedure codes, and other clinical information.

3.2.2 Data Cleaning

The original dataset for our study included 1,385,364 records of patient claims that were associated with beneficiary IDs that had been admitted to hospital for at least one case of primary diagnosis of stroke during January 2014 to December 2015. We excluded 1,275,445 records of claims with primary diagnosis other than stroke. Out of the remaining 109,919 records of claims hospitalized with a primary diagnosis of stroke, additional 35,494 records were removed (not admitted to a short term acute care hospital: 22,777, deceased/expired during hospitalization: 6,519, patients discharged to other locations: 6,198). Of the remaining 74,425 records, 31,625 (42.5%) corresponded to home discharge and 42,800 (57.5%) corresponded to facility discharge (Table 3.1).

3.2.3 Feature Choices

We grouped age into three categories: 18 to 64 years, 65 to 74 years, and 75 years and older. Stroke types were pooled into three different categories: ischemic, meningeal hemorrhage, and intracerebral hemorrhage. We included diabetes, high cholesterol, obesity,

hypertension, atrial fibrillation, other atrial disease, chronic kidney disease, heart disease, peripheral arterial disease, other vascular diseases, prior stroke or TIA, acute heart attack, sleep habits, alcohol habits, drug habits, smoking, family history, depression, and other diagnosis as comorbidities or other possible risk factors. Sources of admission were grouped into five different categories: non-healthcare facility (physician's referral); clinic referral; transfer from a hospital; transfer from a skilled nursing facility (SNF); other facilities. Primary health insurance was divided into medicaid or medicare, private insurance, or other insurances. Hospital discharge disposition status was coded as home discharge when patients were discharged to home with or without home health (HH) care services and defined as facility discharge when patients were discharged to healthcare facilities such as an SNF, an inpatient rehabilitation facility (IRF), and another short-term general hospital for inpatient care.

3.2.4 Statistical Analysis

By dividing the age into three different age groups, all the features become categorical. The Pearson's Chi-square test was used to determine the independency of the features. Based on the result of the Chi-square test, no associations were found between the discharge status and different groups within each feature, considering a significant level of 0.05 (Table 3.1). General collinearity test was performed to the total cohort and no strong collinearity was observed between the different features. Based on the statistical analysis, a multivariate logistic regression model was developed; odds ratios and unadjusted odds ratios as well as their corresponding 95% confidence intervals and coefficients (betas) with significant level of 0.05 were generated to examine the discharge status in the training cohort. Based on the values of the coefficients, different risk factors were evaluated and coded for further analysis.

Table 3.1 Demographic and clinical characteristics of stroke patients

Patient characteristics	Home discharge N = 31,625 (%)	Facility discharge N = 42,800 (%)	P value
Sex			<0.0001
Male	16,038 (50.7)	18,728 (43.8)	
Female	15,587 (49.3)	24,072 (56.2)	
Age			<0.0001
18-64	3,636 (11.5)	3,439 (8.0)	
65-74	10,570 (33.4)	10,211 (23.9)	
$\geq$ 75	17,419 (55.1)	29,150 (68.1)	
Race			<0.0001
White	24,952 (78.9)	33,604 (78.5)	
Black	4,565 (14.4)	6,710 (15.7)	
Other	2,108 (6.7)	2,486 (5.8)	
Stroke Type			<0.0001
Ischemic	28,708 (90.8)	36,855 (86.1)	
Meningeal hemorrhage	1,215 (3.8)	1,698 (3.9)	
Intracerebral hemorrhage	1,702 (5.4)	4,247 (10.0)	
Comorbidity			<0.0001
Diabetes	10,300 (32.6)	14,569 (34.0)	
High cholesterol	17,300 (54.7)	22,333 (52.2)	
Obesity	3,078 (9.7)	4,018 (9.4)	
Hypertension	19,223 (60.8)	25,494 (59.6)	
Atrial fibrillation	9,134 (28.9)	15,551 (36.3)	
Other atrial disease	3,134 (9.9)	4,277 (10.0)	
Chronic kidney disease	5,372 (16.9)	8,626 (20.2)	
Heart disease	12,599 (39.8)	19,171 (44.8)	
Peripheral arterial disease	1,904 (6.0)	2,749 (6.4)	
Other vascular	861 (2.7)	1,121 (2.6)	
TIA	9,104 (28.8)	13,239 (30.9)	
Acute heart attack	311 (1)	1,002 (2.3)	
Sleep habit	907 (2.9)	1,134 (2.6)	
Alcohol habit	936 (3.0)	1,354 (3.2)	
Drug habit	463 (1.5)	549 (1.3)	
Smoking	10,405 (32.9)	11,717 (27.4)	
Family history	2,373 (7.5)	2,351 (5.5)	
Depression	194 (0.6)	326 (0.8)	
Other diagnosis	819 (2.5)	1,572 (3.7)	
Source of Admission			<0.0001
Non-healthcare facility	28,641 (90.6)	36,911 (86.2)	
Clinic referral	1,172 (3.7)	1,443 (3.4)	
Transfer from a hospital	1,437 (4.5)	2,414 (5.6)	
Transfer from a SNF	138 (0.4)	1,463 (3.4)	
Other	237 (0.8)	569 (1.4)	
Type of Insurance			<0.0001
Medicare or Medicaid	30,585 (96.7)	42,129 (98.4)	
Private Insurance	833 (2.6)	534 (1.2)	
Other	207 (0.7)	137 (0.4)	

3.2.5 Part A: Obtaining Risk Scores through Logistic Regression

Based on cohorts selected above, logistic regression was performed to estimate odds ratios (ORs) of patient characteristics associated with facility discharge. Both unadjusted and adjusted ORs with 95% confidence intervals were considered. After that, coefficients (beta) from the multivariate logistic regression model were utilized to derive risk scores [26, 27, 46]. A total risk score was calculated for each patient by taking the sum of corresponding risk scores (see example in Table 3.2). After the logistic function, the predicted probability of facility discharge for each total risk score was presented and compared against the observed counterpart. Lastly, a predictive tool was made by using the total risk score to predict the hospital discharge disposition status of each patient with a primary diagnosis of stroke.

Table 3.2 Total risk score calculation

Beneficiary ID	Discharge Status	Gender	Age	Stroke Type	...	Total Risk Score
A	1	1	3	5	...	11
B	0	0	3	1	...	4
C	1	0	1	0	...	2

3.2.6 Part B: 30-Day Readmission Analysis

After the total risk score was calculated for each patient, the total risk score was converted into a predicted discharged disposition status ($\hat{y}$), to be compared with the actual discharge disposition status (y) for the readmission analysis. Based on the probability of facility discharge for a given total risk score (Fig 3.1), we established a threshold value to assign the value 'home discharge' for total risk scores that are lower than the threshold value, and 'facility discharge' for the scores that are greater than or equal to the threshold value (Table 3.3).

After the conversion step, we separated patients by their discharge disposition status (home or facility) and from there, we further broke down the data into four cases: 1- actual

Table 3.3 Total risk score conversion

Beneficiary ID	Total Risk Score	Actual Discharge Status (y)	(Threshold = 9) Predicted Discharge Status ($\hat{y}$)
A	10	facility	facility
B	7	facility	home
C	14	facility	facility
D	4	home	home
E	9	facility	facility

discharge status is home and predicted discharge status is home, 2- actual discharge status is home and predicted discharge status is facility, 3- actual discharge status is facility and predicted discharge status is home, 4- actual discharge status is facility and predicted discharge status is facility. All four cases were tested to see if the patients returned to hospital within 30 days. A 30-day search window was applied for January 2014 to eliminate claims that were from before 2014. Furthermore, we removed any claims that were recorded after December 1st, 2015 to select cohorts strictly from 2014-2015. After removing data through a searching window, the dataset for our investigation included 66,172 stroke patients with unique beneficiary IDs.

3.3 Results

3.3.1 Results on Analysis Part A

Based on both unadjusted and adjusted ORs, patient characteristics such as female sex; ages 75 years and older; black race; meningeal hemorrhage or intracerebral hemorrhage; presence of diabetes, hypertension, atrial fibrillation, chronic kidney disease, heart disease, acute heart attack, alcohol habit, depression, or other diagnoses; transfer from a hospital, transfer from an SNF, or other were associated with an increased risk of having a facility discharge (Table 3.4).

The range of the calculated risk scores for patient characteristics was from -3 to 13 (Table 3.5). The range of the calculated total risk score for a given patient was from -7 to 29.

The predicted probability of facility discharge increased with the increase in total risk score (Fig. 3.1), which indicates that a patient with a higher total risk score had a higher chance of being discharged to a healthcare facility.

3.3.2 Results on Analysis Part B

Out of 66,172 unique stroke patients who were being tested for 30-day readmission analysis, 28,789 (43.5%) patients were related to home discharge and the other 37,383 (46.5%) patients corresponded to facility discharge. For the case where the actual discharge status is a home and predicted discharge is a facility (n=1,236), 186 (15%) patients were readmitted within 30 days. For the case where both the actual and predicted discharge status are home (n=27,553), 2,640 (9.5%) patients were readmitted within 30 days. For the case where actual discharge status is facility and predicted discharge status is facility (n=4,691), 856 (18.2%) patients were readmitted within 30 days. Lastly, for the case where actual discharge status is a facility and predicted discharge is home (n=32,692), 4,450 (13.6%) patients were readmitted within 30 days (Fig. 3.2).

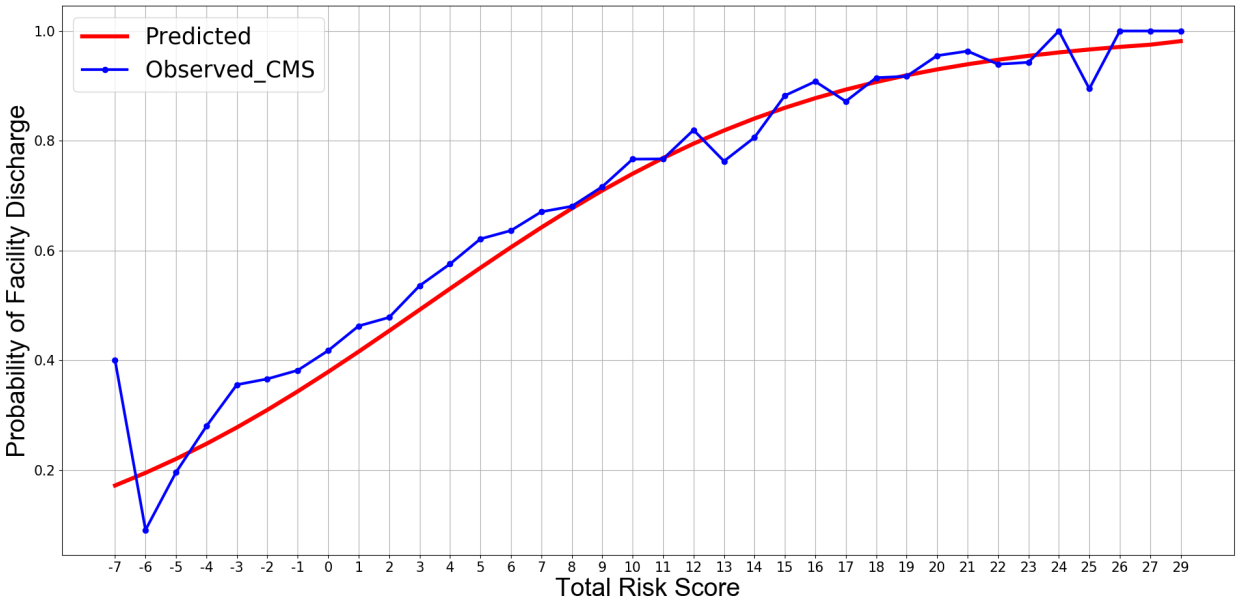


Figure 3.1 Predicted and observed probabilities of facility discharge for each total risk score.

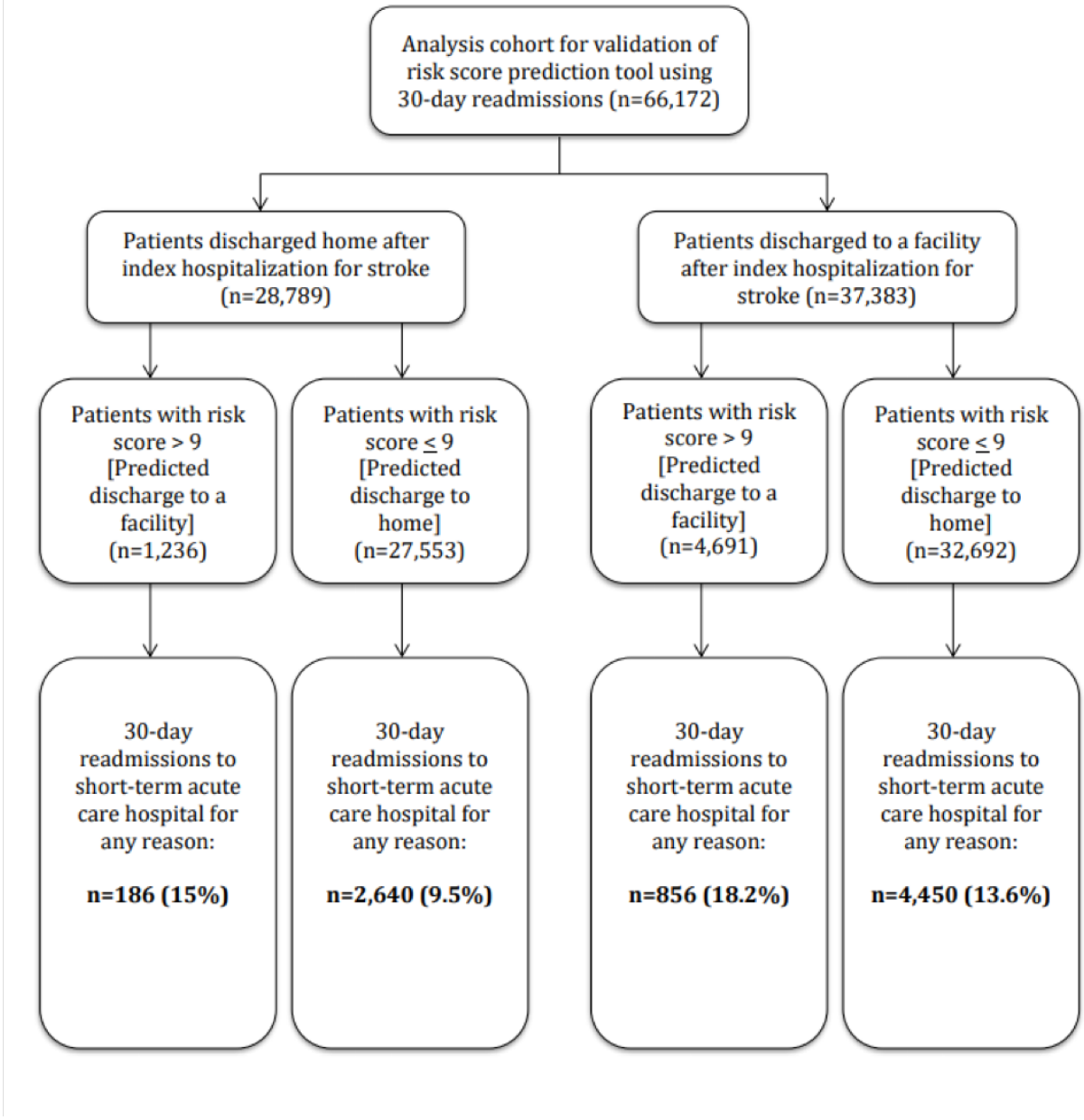


Figure 3.2 Flowchart of validation of risk score prediction tool using 30-day readmission

Table 3.4 Odds Ratios of patient characteristics associated with facility discharge

Patient characteristics	Unadjusted OR	Adjusted OR	β
Sex			
Male	1.00 (Ref.)	1.00 (Ref.)	0
Female	1.32 (1.28-1.36)	1.25 (1.21-1.29)	0.2245
Age			
18-64	1.00 (Ref.)	1.00 (Ref.)	0
65-74	1.02 (0.97-1.08)	1.07 (1.01-1.13)	0.0692
$\geq$ 75	1.77 (1.68-1.86)	1.71 (1.62-1.81)	0.5386
Race			
White	1.00 (Ref.)	1.00 (Ref.)	0
Black	1.09 (1.05-1.14)	1.21 (1.16-1.27)	0.1924
Other	0.88 (0.82-0.93)	0.91 (0.86-0.97)	-0.0942
Stroke Type			
Ischemic	1.00 (Ref.)	1.00 (Ref.)	0
Meningeal hemorrhage	1.09 (1.01-1.17)	1.13 (1.05-1.22)	0.1239
Intracerebral hemorrhage	1.94 (1.83-2.06)	2.02 (1.90-2.14)	0.7020
Comorbidity			
Diabetes	1.07 (1.03-1.10)	1.15 (1.11-1.19)	0.1396
High cholesterol	0.89 (0.86-0.92)	0.92 (0.89-0.95)	-0.0806
Obesity	0.96 (0.92-1.01)	1.07 (1.01-1.12)	0.0647
Hypertension	1.08 (1.04-1.12)	1.05 (1.01-1.09)	0.0476
Atrial fibrillation	1.35 (1.31-1.39)	1.27 (1.23-1.31)	0.2386
Other atrial disease	1.01 (0.97-1.07)	1.06 (1.01-1.12)	0.0605
Chronic kidney disease	1.23 (1.17-1.29)	1.20 (1.14-1.26)	0.1791
Heart disease	1.13 (1.09-1.17)	1.11 (1.08-1.15)	0.1116
Peripheral arterial disease	1.04 (0.98-1.10)	1.04 (0.98-1.11)	0.0423
Other vascular	0.95 (0.87-1.04)	0.99 (0.90-1.09)	-0.0108
TIA	1.09 (1.06-1.13)	1.09 (1.05-1.12)	0.0822
Acute heart attack	2.22 (1.94-2.52)	2.24 (1.97-2.56)	0.8075
Sleep habit	0.95 (0.86-1.03)	0.97 (0.89-1.07)	-0.0277
Alcohol habit	1.21 (1.11-1.32)	1.44 (1.32-1.57)	0.3653
Drug habit	0.95 (0.84-1.08)	1.17 (1.02-1.33)	0.1585
Smoking	0.78 (0.76-0.81)	0.88 (0.85-0.91)	-0.1247
Family history	0.74 (0.70-0.79)	0.76 (0.72-0.81)	-0.2669
Depression	1.27 (1.01-1.52)	1.30 (1.08-1.57)	0.2640
Other diagnosis	1.43 (1.32-1.57)	1.42 (1.30-1.55)	0.3526
Source of Admission			
Non-healthcare facility	1.00 (Ref.)	1.00 (Ref.)	0
Clinic referral	0.95 (0.88-1.03)	0.97 (0.89-1.05)	-0.0342
Transfer from a hospital	1.30 (1.22-1.39)	1.24 (1.16-1.33)	0.2169
Transfer from a SNF	8.22 (6.90-9.80)	7.00 (5.87-8.35)	1.9461
Other	1.86 (1.60-2.17)	1.79 (1.53-2.09)	0.5820
Type of Insurance			
Medicare or Medicaid	1.00 (Ref.)	1.00 (Ref.)	0
Private Insurance	0.47 (0.42-0.52)	0.62 (0.55-0.69)	-0.4821
Other	0.48 (0.39-0.60)	0.69 (0.55-0.86)	-0.3703

Table 3.5 Risk scores of patient characteristics

Patient characteristics	Risk Score
Sex	
Male	0
Female	1
Age	
18-64	0
65-74	1
$\geq$ 75	3
Race	
White	0
Black	1
Other	-1
Stroke Type	
Ischemic	0
Meningeal hemorrhage	1
Intracerebral hemorrhage	5
Comorbidity	
Diabetes	1
High cholesterol	-1
Obesity	0
Hypertension	0
Atrial fibrillation	2
Other atrial disease	0
Chronic kidney disease	1
Heart disease	1
Peripheral arterial disease	0
Other vascular	0
TIA	1
Acute heart attack	5
Sleep habit	0
Alcohol habit	2
Drug habit	1
Smoking	-1
Family history	-2
Depression	2
Other diagnosis	2
Source of Admission	
Non-healthcare facility	0
Clinic referral	0
Transfer from a hospital	1
Transfer from a SNF	13
Other	4
Type of Insurance	
Medicare or Medicaid	0
Private Insurance	-3
Other	-2

3.4 Discussion

This study validated a discharge disposition predictive tool using integer-based risk scores for patients at index hospitalization for stroke as well as its utility in reducing readmission rates. Of the patients who were discharged to home, the algorithm predicted 95.7% of them to have that discharge disposition. In the readmission analysis, the scenario of predicted discharge to home and actual discharge to home only had a readmission rate of 9.5%, which is well below the usual readmission rate for patients post-stroke [55–57].

Creating predictive tools to better match patients with an appropriate discharge destination may decrease the transition time from admission to discharge, whether to home or facility. Clinicians may be able to better identify high-risk patients and initiate more complex discharge planning early in a patient's length of stay. Additionally, unnecessary readmissions may be prevented by matching a patient more accurately with their appropriate discharge location. Improved matching may result in fewer complications and better functional recovery. These predictive tools can be simple and quick to use and may decrease the length of stay and readmissions, thus reducing costs. The top five risk scores found to be predictive of discharge disposition were admission from an SNF, acute myocardial infarction, intracerebral hemorrhage, admission from 'other' source, and an age of 75 or older. Myocardial infarction and age of 75 or older are risk factors for stroke [59] and are common indicators for a more complex medical management [60]. Older patients are likely to have more comorbidities and less support at home compared to younger patients and may require further medical care and monitoring at a facility. Intracerebral hemorrhage expectedly has a high-risk score as it is considered more severe than ischemic stroke or transient ischemic attack as evidenced by its correlation with an increase in mortality [32].

We used readmission rates as an indicator of prediction tool quality due to the significance of this metric for hospital administrators and clinicians alike. Relevant literature encourages hospitals to take measures to identify high-risk patients for readmission and

determine appropriate discharge disposition and follow-up services in order to reduce readmission rates [55,57]. Readmission rates are a rising concern for both hospital administrators and clinicians alike. This common ground makes lowering unnecessary readmissions a high priority focus amongst the healthcare team. Much research has explored predictors of readmission in conditions such as type 2 diabetes mellitus [53], stroke [55], heart failure, acute myocardial infarction, pneumonia, and chronic obstructive pulmonary disease [52]. Some of these admissions are often questioned as potentially preventable, and hospital staffs are encouraged to identify high-risk patients and intervene accordingly [55, 56]. This study contributed to the current literature by validating a predictive discharge disposition tool with readmission rates.

The predictive tool created in this study predicted home discharge with extremely high occurrence. This may have been due to the high scores attributed to admission source versus comorbidities. While it is clinically apparent that patients receiving medical management immediately prior to stroke are likely to require continued management after their short-term acute care hospital stay, this score may have diminished the effect of other variables that help distinguish the significance of factors such as comorbidities and lifestyle behaviors.

There were several limitations to this study. The findings are limited to two years of Medicare beneficiaries and may not be generalizable to all patients post-stroke. Some patients may have lost insurance coverage after discharge and their readmissions art not recorded in the CMS dataset. In future research, it would be beneficial to exclude those patients from the analysis cohort. Risk scores were calculated based on index hospitalization for stroke; however, we could not know if this was the patient's first stroke or if it was a recurrent stroke with the first stroke occurring prior to our dataset. Patients with recurrent strokes would likely be considered at higher risk for facility discharge, however, this could not be accounted for without a full admission history. When validating the predictive tool via readmission analysis, the threshold to determine when the algorithm would predict facility versus home discharge was arbitrarily set at 75% probability of facility discharge. However,

this threshold could likely be adjusted to allow for a closer match between predicted and actual discharge dispositions. Collaborating with hospital administrators or physicians may allude to a more clinically meaningful threshold that increases confidence in relying on the predictive tool. The top risk score factor was the admission from an SNF, which was dramatically higher than the next highest factor. This score may have shifted the probability curve and resulted in high levels of predicted home discharge for patients admitted from any other source. Clinically, admission from an SNF indicates a patient with high medical management pre-stroke, and discharge back to a facility is assumed to be likely. Existing studies have pointed to decreased outcomes for patients admitted to SNFs in comparison to home [13] or IRF [45]. It is difficult to determine whether the disparity in outcomes is due to the patients' medical complexity or the type, quality, and amount of care received at a SNF. Because of this, admission source may not be an insightful variable that adds to the general clinical reasoning. Eliminating this variable may depress the risk scores and give greater weight toward comorbidities and stroke type. Additionally, details of each patient's characteristics are limited to the amount of detail in their claim. We did not track the role of factors such as functional status, treatments received during the acute care stay, or patient and family preference in determining discharge status. These factors may provide deeper insight into a patient's profile. Lastly, while readmission rates are a well-accepted measure of the quality of care, we are unable to distinguish if any given readmission was due to inappropriate discharge planning or poor quality of care along the patient's journey.

3.5 Conclusion

In this study, we developed a discharge disposition prediction tool for use after index hospitalization post-stroke. We utilized a probabilistic model (logistic regression) to assess the relationship between the outcome variable (discharge status) and its predictors (patient characteristics). Regression coefficients were converted into risk scores to determine the

probability of facility discharge using our probabilistic model. The advantage of using this model is the ability to generate both positive and negative scores. The discharge outcome was efficiently calculated by assigning risk scores to each patient. Many patients and hospital-related factors affect the discharge disposition, making it a complex decision-making process. Prediction tools are helpful to guide clinicians and hospital administrators as they seek ways to improve the quality of care and reduce preventable readmissions through efficient and appropriate discharge planning.

CHAPTER 4
MACHINE LEARNING BASED SMART HEALTH

4.1 Introduction

In the era of big data, as large-scale data become more easily accessible, researchers have been taking advantages of sophisticated algorithms to tackle research problems in more efficient way. The rapid advances and recent improvements in machine learning algorithm has aided researchers in many fields (i.e., healthcare, energy, transportation, and etc.) by providing robust and efficient predictive techniques. Some of these machine learning algorithms have shown their increased effectiveness in solving complex problems when compared to the conventional statistical methods. Despite their promising results and remarkable accuracy, their convoluted and non-linear structure induces a major issue in the model transparency [61]. In linear models such as linear regression and logistic regression, the relationship between a dependent variable and independent variables are apparent [62], and their linear structure allows an interpretation of model parameters [63]. However, in machine learning algorithms, it is difficult to investigate the effect of the information that the input data provides to the final decision. Thus, these machine learning models are often referred to as "black-box models" [61], and do not allow for the interpretation of model parameters. The perception of the model's behavior and the reason behind predictions is essential when the model is used for making critical decisions [64]. When machine learning models are used for medical diagnosis [65] or other outcome related predictions [66], the predictions should not be trusted easily without any inference of the model, as the consequences could negatively impact human beings [64].

4.2 Contribution

In this section, interpretable machine learning models are used to predict hospital discharge disposition of stroke patients using five years of data from the Tennessee Department of Health. We have investigated three machine learning algorithms (i.e., Random Forest, AdaBoost, and MLP) and one linear model (i.e., Logistic Regression) as a comparison. To interpret the results, we have used LIME to provide comprehensive interpretation of predictions made by these black-box models. Our main contributions are outlined as follows:

- The use of black-box machine learning algorithms to predict hospital discharge disposition of stroke patients in Tennessee.
- The use of LIME to interpret the results of the black-box models and comparing it with the previously used inherently interpretable model.

4.3 Related Work

The majority of previous studies in the medical domain have relied heavily on multivariate logistic regression to predict the outcome under investigation [26, 27, 67]. Our previous study [46] also utilized logistic regression to predict hospital discharge disposition of stroke patients. Based on the result of logistic regression, risk scores for predicting facility discharge of stroke patients were developed. Using the risk scores, we were able to discover significant risk factors associated with the facility discharge (Table 4.1). However, we were not able to demonstrate significant predictive capabilities due to the limitation of the linear model.

A machine learning approach in making medical diagnosis was studied by [68]. The authors of this study surveyed different investigations that utilize machine learning algorithms in cancer prediction and prognosis. They found that the majority of published studies were able to substantially improve the accuracy of predicting cancer susceptibility, recurrence, and mortality. Regardless of its effectiveness, the authors pointed out the difficulties in

Table 4.1 Top five feature scores associated with facility discharge obtained from logistic regression

Rank	Feature Name
1	Intracerebral hemorrhage
2	Subarachnoid hemorrhage
3	Hypertension
4	Ages 75 years and older
5	Arrhythmia

understanding the complex structure of machine learning algorithms. Nevertheless, the discriminating power of machine learning algorithms are generally much better than linear models, which is the reason behind their popularity in domains where classification performance is valued greater than model interpretation [63]. Interpretability has higher priority in the medical field, and hence, we used logistic regression in our previous study [46] due to its ability to explain the predictions based on the model parameters. Unfortunately, the performance of logistic regression was unsatisfactory for its inability to capture non-linear relationships. One solution to this paradox is using black-box models with the assistance of another external interpretation method. Moreover, a model agnostic method is required to provide the freedom of choosing any model, while producing scores for the importance of different features used to get the final prediction. This applies typically on Local Interpretable Model-agnostic Explanations (LIME).

LIME is a model agnostic machine learning interpretation method that explains the individual predictions by approximating them locally using a linear model [64]. There are other interpretation methods that study the effects of more than one feature in making the final predictions [69], and calculate the combined score for multiple features. However, the main drawback of this method is that it is structure dependent, and hence adjustments are required to be applied to different machine learning architectures.

4.4 Methods

4.4.1 Cohort Selection

We used data from the Hospital Discharge Data System managed by the Tennessee Department of Health. The objective of the Hospital Discharge Data System is to obtain and summarize hospital claims data to analyze charges for types of services that are related. The dataset included all of the records of hospitalized patients with the principal diagnosis of stroke(*International Classification of Diseases, Ninth Revision* codes 430, 431, 433, 434, 436). This data contained the following information:

- Patient Demographics: sex, age, and race
- Stroke Type (ICD-9): principal diagnosis code and other diagnosis codes related to stroke
- Source of Admission
- Insurance Information: primary and secondary payer classes
- Discharge Disposition Status

4.4.2 Data Cleaning

The raw data for our study included 139,706 records of patient claims that were associated with the primary diagnosis of stroke from 2010 to 2014. We have removed 12,125 records of claims (missing data: 1,151; expired: 6,855; discharged to hospice: 3,185; discontinued care and court: 934).

4.4.3 Feature Selection

Age was given three categories: 18 to 64 years, 65 to 74 years, and 75 years and older. Stroke types were pooled into three categories: ischemic, SAH, and ICH. We selected diabetes mellitus, heart disease, hypertension, peripheral arterial disease, chronic kidney disease, hyperlipidemia, arrhythmia, and depression as comorbid conditions. Sources of admission to the hospital were grouped into home or a non-healthcare facility, clinic or physician's office, or another hospital. Primary and secondary health insurances were categorized into private insurance, Medicaid, managed Medicare, and Medicare fee-for-service. Our target variable, discharge disposition status was labeled as "discharged to home" when patients were discharged home with or without home health care services and as "discharged to facility" when patients were discharged to healthcare facilities such as an skilled nursing facility (SNF), an intermediate care facility, inpatient rehabilitation facility (IRF), and another short-term general hospital for inpatient care. Variables that were used in the study are shown in Table 4.2.

Table 4.2 Data Variables Considered in Study

Data Variables	Description
Discharge Disposition Status	Home / Facility
Sex	Male / Female
Age	18 to 64 / 65 to 74 / 75 above
Race	White / Black / Other
Stroke Type	Ischemic / SAH / ICH
Comorbidities	Diabetes, Heart disease, ... , Depression
Source of Admission	Home or Non-healthcare / Clinic / Another hospital
Primary Insurance	Private / Medicaid / Medicare1 / Medicare2
Secondary Insurance	Private / Medicaid / Medicare1 / Medicare2

As the main objective of the study is to have the interpretation capability in black-box models while maintaining their dominating predictive ability within our application, we introduce one baseline method (Logistic Regression) and three black-box models (Random Forest, AdaBoost, and MLP) to achieve our goal. First, we fit the three models on the

hospital discharge data to achieve better accuracy compared with the baseline model. Then we interpret the results using LIME.

4.4.4 Explanation

LIME method explains the outputs of a classifier by approximating them locally using a linear model. First, a set of new samples is generated around that instance. Then, these instances are applied to the black-box model in order to calculate their prediction probabilities. This results in a mini-dataset of new instances. The next step is to use this mini-dataset to fit a linear model and consider its coefficients as the importance scores of the model features. The process explained above generates explanation for only one sample. In order to understand the general behaviour of the model, we have to construct a method to identify the most representative samples. LIME suggested a sub-modular pickup algorithm to perform this step [64]. This algorithm selects several instances from the dataset so that their explanations are diverse and representative to the model's features. The number of the selected instances is determined by the user and it represents how many samples we can look into them to understand the model and not get confused into the details. In this study, 20 samples have been selected by the sub-modular algorithm for the four models. Then the final 80 samples are used to generate explanations from all the models. This allows for comparison between the scores of the four models and the sub-modular algorithm samples for each model.

4.4.5 Baseline Model: Logistic Regression

Logistic regression is a well-known linear model that has been used extensively in solving classification problems (mostly with dichotomous dependent variables) for its simplicity, interpretability, and the ability to predict with probability estimation [46, 70]. An equation

for multivariate logistic regression with k independent variables is shown in Equation 4.1. As demonstrated in our previous study [46], odds ratios (ORs) of features (X_i) associated with facility discharge (Y) were estimated based on the result of multivariate logistic regression. After obtaining the ORs, coefficients (β_i) from the ORs were used to derive risk scores, which can be used to find the total risk score. Examples of this score calculation are provided in Table 3.2 for three patients, labeled A, B, and C.

$$Pr(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)} \tag{4.1}$$

4.4.6 Black-box Models with LIME

Three black-box models have been applied on the data: Random Forests, Random Forests with AdaBoost classifier, and Multi-layer Perceptron (MLP).

- **Random Forests**: A random forest classifier fits different decision trees on sub-samples from the training set [71].
- **Random Forest with AdaBoost Classifier**: Adaboost classifier is a model in which multiple random forests classifiers are fit on different copies of the training dataset starting from the whole dataset and then focusing only on the mis-classified instances [72].
- **Multi-layer Perceptron (MLP)**: MLP is a type of feed-forward artificial neural network that is made up of three or more layers:the input layer, hidden layers, and the output layer. In MLP, data is moved from input layer to output layer in one direction through a backpropagation learning algorithm. MLPs are widely used for both estimation and classification problems. For our study, we have used the Rectifier activation function for the the hidden layer. we used a MLP with two layers of 20, 8 neurons respectively with ReLU activation function followed by sigmoid function.

4.5 Results

Table 4.3 Index of the Features

Term	Long Name	Term	Long Name
f1	Gender	f9	Has Chronic Kidney
f2	Age	f10	Has Hyperlipidemia
f3	Race	f11	Has Arrhythmia
f4	Stroke Type	f12	Has Depression
f5	Has Diabetes	f13	Source of Admission
f6	Has Heart Disease	f14	Primary Payer Class
f7	Has Hypertension	f15	Secondary Payer Class
f8	Has Peripheral Arterial	–	–

Out of 127,581 records remaining after data cleaning, 86,114 (65.5%) were related to home discharge and 41,467 (32.5%) were related to facility discharge (Table 4.5). For performance evaluation, data was selected from 2010 to 2013 (101,223 records) as the training set and the remaining 26,358 records (2014) as the testing set.The performances of both training and testing prediction accuracy is shown in table 4.4. In terms of testing performance, the random forest model achieved 69% accuracy. The performances of the logistic regression model and the AdaBoost model were slightly better with prediction accuracy of 70%. The MLP model achieved 71% prediction accuracy and performed the best out of the four models that were considered.

Table 4.4 Performance of hospital discharge disposition classifications

Model	Train Acc	Test Acc	Precision	Recall	F1
Logistic Regression	70%	70%	60	26	58
Random Forest	78%	69%	56	36	61
AdaBoost	71%	70%	60	30	60
MLP	72%	71%	64	26	59

Table 4.5 Demographic and clinical characteristics of stroke patients

Characteristics	Home Discharge (n = 86,114)	Facility Discharge (n = 41,467)	P value
Sex			<0.0001
Men	43,955 (51.0%)	18,708 (45.1%)	
Women	42,159 (49.0%)	22,759 (54.9%)	
Age			<0.0001
18-64	36,136 (41.9%)	13,604 (32.8%)	
65-74	25,673 (29.8%)	9,896 (23.9%)	
≥75	24,305 (28.3%)	17,967 (43.3%)	
Race			<0.0001
White	71,469 (82.9%)	33,114 (79.9%)	
Black	11,533 (13.4%)	7,012 (16.9%)	
Other	3,112 (3.7%)	1,341 (3.2%)	
Stroke Type			<0.0001
Ischemic	78,774 (91.5%)	34,143 (82.3%)	
Subarachnoid hemorrhage	3,184 (3.7%)	2,383 (5.8%)	
Intracerebral hemorrhage	4,156 (4.8%)	4,941 (11.9%)	
Comorbidity			
Diabetes	21,353 (24.8%)	14,357 (34.6%)	<0.0001
Heart disease	30,237 (35.1%)	21,205 (51.1%)	<0.0001
Hypertension	48,877 (56.8%)	32,055 (77.3%)	<0.0001
Peripheral arterial disease	5,831 (6.8%)	2,120 (5.1%)	<0.0001
Chronic kidney disease	6,004 (7.0%)	5,322 (12.8%)	<0.0001
Hyperlipidemia	27,892 (32.4%)	15,006 (36.2%)	<0.0001
Arrhythmia	10,150 (11.8%)	10,766 (25.9%)	<0.0001
Depression	4,730 (5.5%)	3,486 (8.4%)	<0.0001
Source of Admission			<0.0001
Non-healthcare facility	56,752 (65.9%)	30,788 (74.2%)	
Clinic or physician's office	19,134 (22.2%)	1,696 (4.1%)	
Transfer from a hospital	6,014 (6.9%)	4,544 (10.9%)	
Others	4,214 (5.0%)	4,439 (10.8%)	
Primary Payer Class			<0.0001
Medicare (Not managed)	40,441 (46.9%)	23,645 (57.0%)	
Medicare (Managed)	14,172 (16.5%)	6,740 (16.3%)	
Medicaid	633 (0.7%)	262 (0.6%)	
Private Insurance	23,021 (26.7%)	7,586 (18.3%)	
Others	7,847 (9.2%)	3,234 (7.8%)	
Secondary Payer Class			<0.0001
Medicare (Not managed)	6,327 (7.3%)	3,042 (7.3%)	
Medicare (Managed)	2,143 (2.5%)	1,162 (2.8%)	
Medicaid	5,725 (6.6%)	4,302 (10.4%)	
Private Insurance	24,133 (28.0%)	12,379 (29.9%)	
Others	47,786 (55.6%)	20,582 (49.6%)	

For interpretation, first, the sub-modular pick algorithm is used to generate 20 representative samples for each model. Then, total of 80 samples are used to generate explanations from all the models. The average of the absolute values is calculated to show the relative importance of all the used features from the perspective of the four models including the inherently interpretable model (logistic regression). After that, interpretation scores from the proposed models were compared with the risk score that was developed in our previous study [46]. Figure 4.1 shows the normalized scores for each features to compare the relative importance between the methods. The description of features are explained in Table 4.3. Figure 4.2 shows the most important features for one sample considered in LIME.

The objective of this study is to interpret black-box models within our application (hospital discharge prediction), so that we can utilize their capability of producing higher accuracy. As we can see from the general trend in Figure 4.1, one can conclude that almost all four models agree on which features are more important. Particularly, age, diabetes, hypertension, source of admission, and primary payer class were almost chosen by all models to be important in predicting the relative importance of features. On the other hand, features like heart disease, chronic kidney, depression, and secondary payer class were chosen to be less important features.

4.6 Conclusion

In this section, we demonstrated a machine learning approach to predict hospital discharge disposition and we were able to verify the effectiveness of LIME in providing explanations for prediction results. Our results aligned with our previous study [46] (which was supported by domain experts) in determining the most effective risk factors related to facility discharge. The performance of these algorithms were confirmed with data from Tennessee Department of Health. We will continue this investigation by exploring other machine learning models and fine-tuning existing models to increase performance.

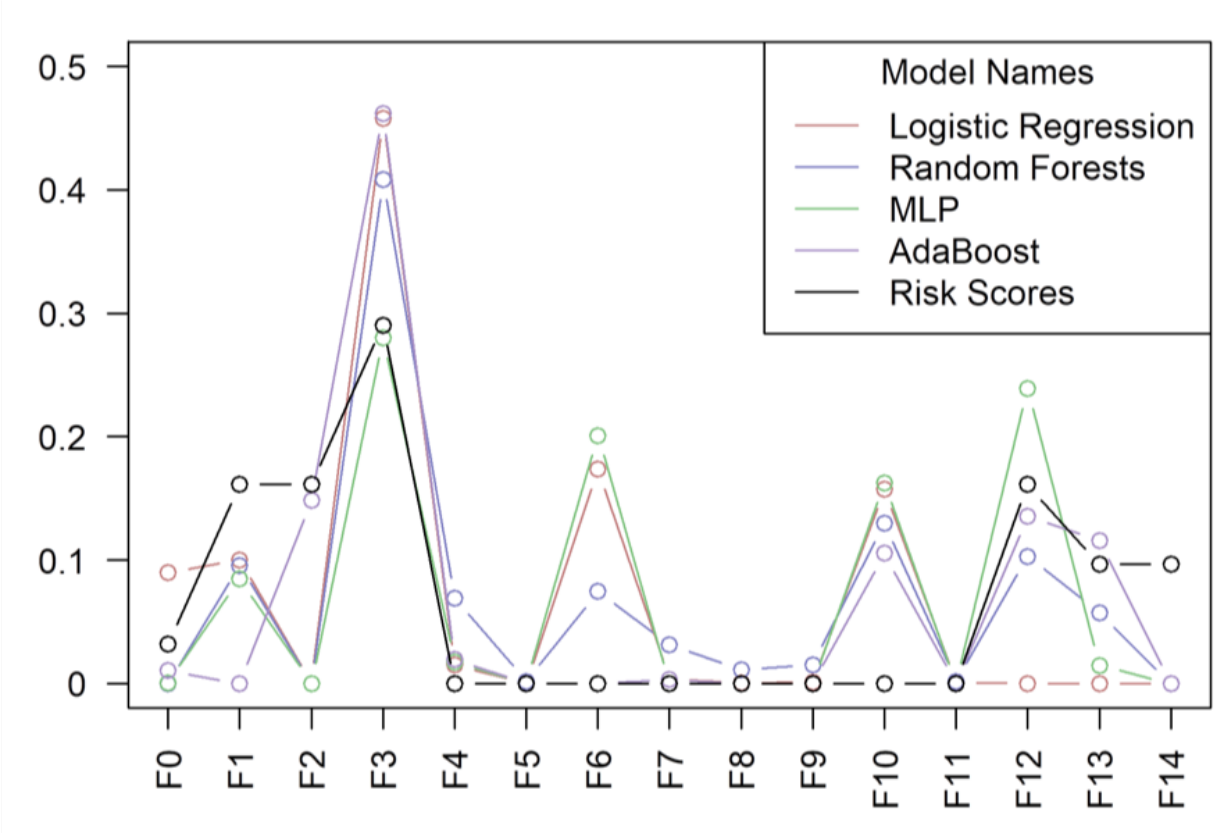


Figure 4.1 Normalized feature scores from the proposed models

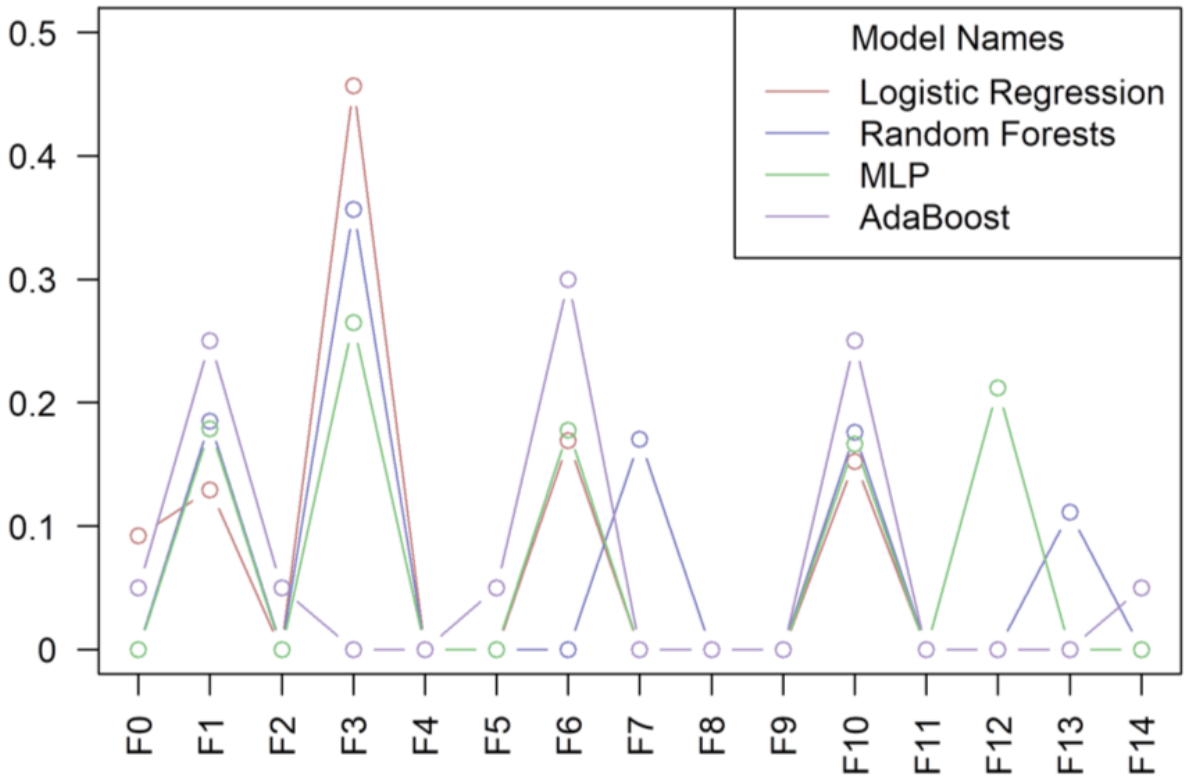


Figure 4.2 Explanations for one sample

CHAPTER 5

The Heavy Lifting Treatment Helper (HeaLTH) Algorithm

5.1 Introduction

One of the astonishing feats of this century is the rapid advancement of medical data and knowledge through clinical studies. With the advancement of computer algorithms and medical devices, we now have access to more data than ever before. This is opening up previously unexplored ways of disease identification and treatment through data science and engineering. Having patient-specific access to data from numerous sensors and data sources allows for intricate and individualized treatment plans to be implemented and advised to patients for illnesses that were once considered incurable. Cancer research in particular is one of the fields that is hoping to take advantage of such novel treatment strategies with a consistent increase in cancer cases. Although the death rate per 100,000 people has decreased since 2010, the total number of deaths from cancer in the U.S. alone has increased from 574,738 in 2010 to 599,099 in 2017 [73], and it is projected to pass 620,000 by 2020 [74]. The number of registered clinical trials has quadrupled from 82,000 studies in 2010 to more than 349,000 in 2020 (reference: clinicaltrials.gov). From these studies, more than 52,000 are currently enrolling patients. However, only one in twenty of cancer patients enroll in clinical trials due to lack of access and complexity of finding the right match for patients [75]. With the expansion of these data sets also comes challenges for clinicians to select the treatment plan that best matches a patient's medical history. Artificial Intelligent (AI) and Machine-learning (ML) algorithms aim to facilitate clinician decision-making by finding similarities in large data sets and combine massive amounts of information from a large pool of patients.

The Heavy Lifting Treatment Helper (HeaLTH) Algorithm proposed here aims to assist clinicians in clinical trial matching for cancer patients using the combination of logical brute-force approach and machine learning algorithms such as agglomerative clustering on clinical trial descriptions.

5.2 Related Works

The process of automatically identifying and clustering trials and eligibility features together based on similarity was performed in [76]. This was accomplished through the construction of a trial-feature matrix comprised of extracted semantic features from the text of the eligibility features for the clinical trials. Through the use of center-based clusters, pairwise similarities were calculated for each clinical trial based on the eligibility features. By using center-based clusters, a single trial was used as the center for each pairwise comparison, allowing for the identification of trials whose similarities to the center trial were no less than 0.9. The team performed their tests on 145,745 clinical trials and extracted a total of 5.5 million semantic features with 459,936 of those features being unique. 8806 center-based clusters were generated, and a sample of those clusters was evaluated using Amazon Mechanical Turk (MTurk) yielding a mean score of 4.331 (on a scale of 1-5).

The team of [77] sought to automate the processes of feature-based indexing, clustering and searching for clinical trials. Their approach was to decompose 80 randomly selected trials for Stage 3 Breast Cancer into a vector of eligibility features organized into a hierarchy. Trials were clustered based on the similarity of their eligibility features. To test their method, the team performed a simulated trial search process by manually selecting features to be used for generating eligibility questions for trial filtering. 1437 distinct eligibility features were extracted, and 80 trials were used. This resulted in 6 clusters which contained trials that took similar patient by patient features, 5 clusters based on disease features, and 2 clusters using mixed features. Additionally, the team demonstrated the utility of named entity recognition

by mapping most features to one or more Unified Medical Language System Concepts.

Similarly, researchers [78] have used Natural language programming to increase clinicians' efficiency in selecting the right clinical trials for pediatric cancer patients. The selected narrative notes from 55 clinical trials from the clinicalTrials.gov and combined that with electronic health records from 215 oncology patients. With automation of the eligibility criteria, they were able to reduce the number of clinical trials matched and saved time for oncologists in choosing the right treatment plan.

5.3 Methods

5.3.1 Data

The data used in this project was provided as part of Oak Ridge National Lab, SMC conference data challenge 2020 which were originally derived from the United State government Clinical Trials website (ClinicalTrials.gov). It consists of 100 cancer patient records (SMC Dataset 2) containing information such as patients' age, gender, therapy history, Performance Status, as well as white blood cell (WBC) count, hemoglobin, platelets, and more (see Table 5.1 for a complete list of variables used in the study). Additionally, six eligibility criteria documents containing the subsets of the clinical trials (SMC Dataset 1) were provided. Each document lists clinical trials pertaining to particular variables seen in Table 5.1, with a total of 1005 trials across all datasets. The eligibility criteria documents contain six factors for clinical trial eligibility presented in Figure 5.1, SMC Dataset 1. These factors are Hemoglobin count, WBC count, Platelets count, HIV, Performance Status, and Prior Therapy. For example, for the WBC factor, the clinical trials have inclusion and exclusion criteria related to a patient's white blood cell count. Additionally, each eligibility file contains seven columns, which can be seen in Table 5.2. Of note is the NCIT column in each eligibility file, which contains a logical statement using c-codes. C-codes are numerical codes that represent medical terminology, e.g., C25150 is age, C12767 is the pelvis. These codes

represent human body parts, basic human information (age and gender), therapy trials, and more. Figure 5.1 shows the flowchart of data sources as well as the detailed steps we took to run our conditional logic and clustering analysis.

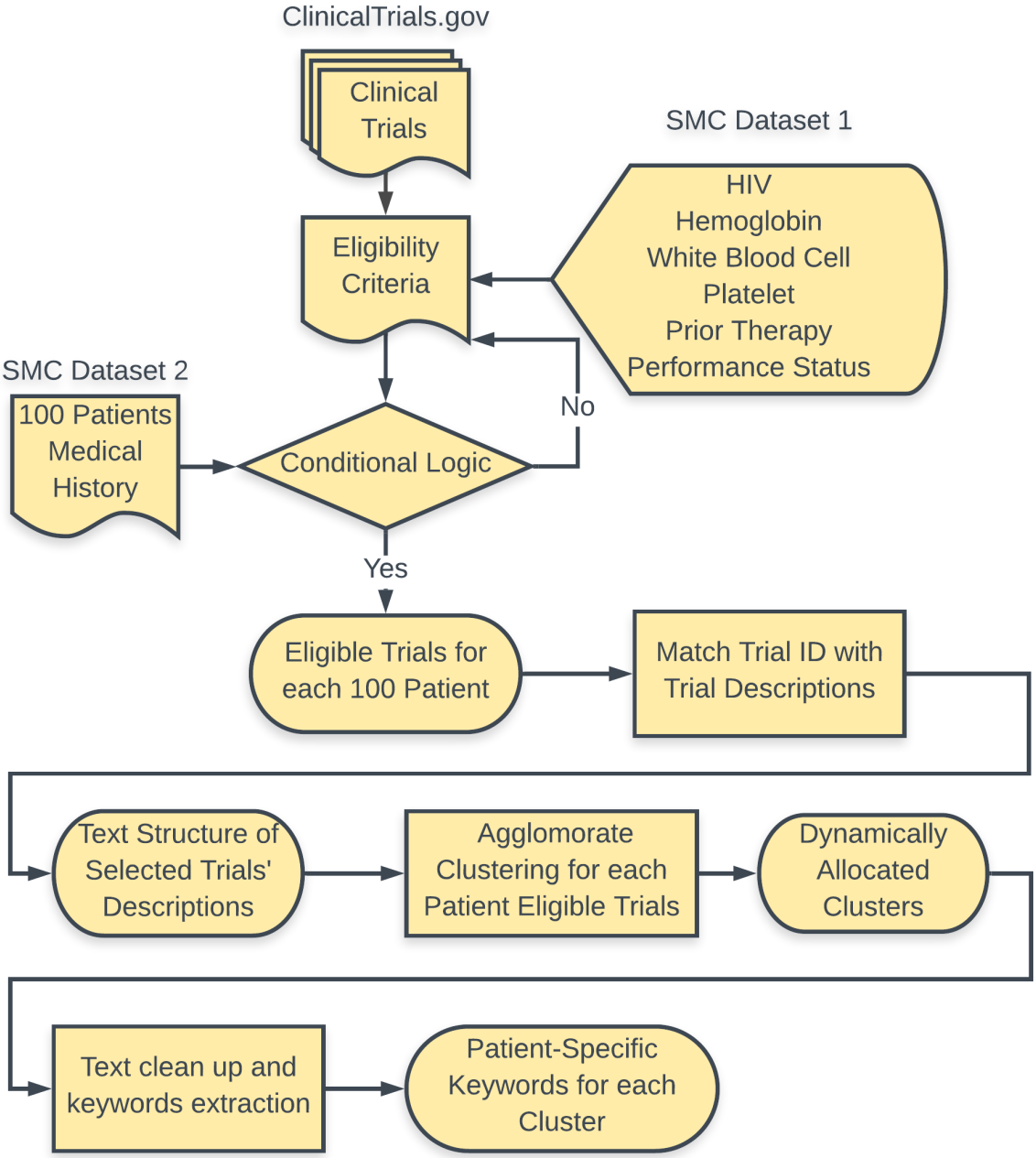


Figure 5.1 Data processing Framework

Table 5.1 Patient Information

Variable	Description
PatientID	numerical value for a patient
Cancer Site (bool)	location of the cancer within the body
Cancer Stage (bool)	stage of the cancer
Treatment History (bool)	prior therapy undergone by patient
Gender	the patient's biological gender
Age	the patient's age
Hemoglobin	patient's hemoglobin count
Platelet	patient's platelet count
White Blood Cell	patient's white blood cell count
Performance Status (bool)	patient's ability to perform daily living activities

Table 5.2 Eligibility File Columns

Variable	Description
NCI_ID/NCT_ID	codes representing different trials
Official Title	the official title of the trial
Inclusion Indicator	include or exclude the patient if they match the criteria
Description	word and logical representation of matching criteria
Text	text version of matching criteria
NCIT	c-code representation of matching criteria

5.3.2 Logical Comparison

To assist in, and act as a baseline for, treatment matching, simple logical operations were performed on the c-codes for each trial in the different eligibility files. For example, in the WBC_Trials dataset, the NCIT column contains several logical statements per trial, such as *C51948>=4000*, which translates to white blood cell count greater than or equal to 4000 per milliliter of blood. The logical code takes the logical statements that accompany each trial in the eligibility file, finds the corresponding information that each c-code represents in a patient's record, and calculates the logic. Any trial that returns a True statement is saved as a potential trial for that patient.

The first step necessary for logical comparison was the cleaning of the NCIT column values, as many entries had a mismatched number of parenthesis, missing c-codes, or blatant syntax errors. Once cleaned, each NCIT conditional statement was read in one at a time

and broken into separate parts. For example, the statement *C51948>=4000* was broken into three segments: *Code: C51948*, *CompOp: >=*, and *Value: 4000*. The code segment for each NCIT conditional was read in and the appropriate patient information was substituted in. So, for the code *C51948*, the patient's white blood cell count was placed in the code's place, and the three segments were combined to create a conditional statement. After the substitution, the statement *C51948>=4000* becomes *X>=4000*, where X represents the current patient's white blood cell count.

This process of patient data substitution was repeated for each portion of a conditional statement, as many trials had many conditional statements for inclusion or exclusion . The output of the logical statement returned a True or False for the whole trial in regards to whether or not the patient met the criteria for inclusion or exclusion.

5.3.3 Clustering

5.3.3.1 Preprocessing

The output of the logical comparison step is merged with the eligibility criteria dataset (e.g., hemoglobin trials, HIV trials, performance status trials, platelets trials, prior therapy trials, WBC trials). From the available columns of the merged dataset, the description, NCTid, and patientID columns are extracted and used for cluster assignment. The primary variable used for the creation of clusters is the "Description" column in the eligibility criteria datasets, while the other variables act as identification factors for the patient(s) and the clinical trials. Natural Language Processing (NLP) techniques were applied to the dataset to pre-process and clean up the text, extract keywords, apply term frequency-inverse document frequency (TFIDF) to get the frequency of those keywords. All the rows with NAN values were also removed from the dataset.

5.3.3.2 Jaccard Similarity

After the pre-processing step, the Jaccard similarity index is calculated to determine the similarities between the two sets of words. Jaccard only takes a unique set of words in each sentence, and the repetition of words does not reduce the similarity index. This is why it is preferred over other similarity measures such as cosine similarity, which takes the length of words of vectors [79]. We have applied lemmatization to reduce the words to the same root words and selected pairwise distance to compute the Jaccard similarity index. If the sets are similar, the similarity index will be equal to 1, otherwise, it will be equal to 0. Equation 5.1 shows how this similarity index is calculated.

$$J(A, B) = \frac{A \cap B}{A \cup B} \tag{5.1}$$

5.3.3.3 Agglomerative Clustering

Agglomerative clustering is a type of hierarchical clustering technique is that well-established in unsupervised machine learning [80]. In agglomerative clustering settings, the dataset is partitioned into singleton nodes and merged one by one with the current pair of mutually closest nodes into a new node until it is left with one last node, which makes up the whole dataset. This clustering method is different from other clustering methods in a way that it measures the inter-cluster dissimilarity and updates that after each step [80,81]. The clustering is applied to the trials which make it past the logical comparison filter. Once clustering is applied, there are N number of clusters that contain X number of possible trials. The number of clusters was selected dynamically depending on the size of trials for each patient. To find the optimal k number of clusters, we have computed the following equation:

$$k = floor(log_2(length(eligible\_trials))) \tag{5.2}$$

Table 5.3 Sample Trial Match Returns

NCI_ID	NCT_ID	Patient_ID
NCI-2009-00336	NCT00392327	1
NCI-2011-00878	NCT00956007	1
...	...	...
NCI-2016-00071	NCT03077451	100
NCI-2016-00787	NCT03030417	100

5.4 Results

The result of the logical comparison step is returned as a list of eligible clinical trials for each patient. The sample trials are presented in Table 5.3. Upon completion of the logical comparison step, the resulting list seen in Table 5.3 has clustering applied on a patient by patient case. Upon the completion of the logical comparison step, the average eligible trials across the 100 patients provided through the data challenge were 283 ±69 from the total 1005 available trials across the six eligibility criteria files. The reduced number of trials for the first 10 patients in our dataset is presented in Figure 5.2.

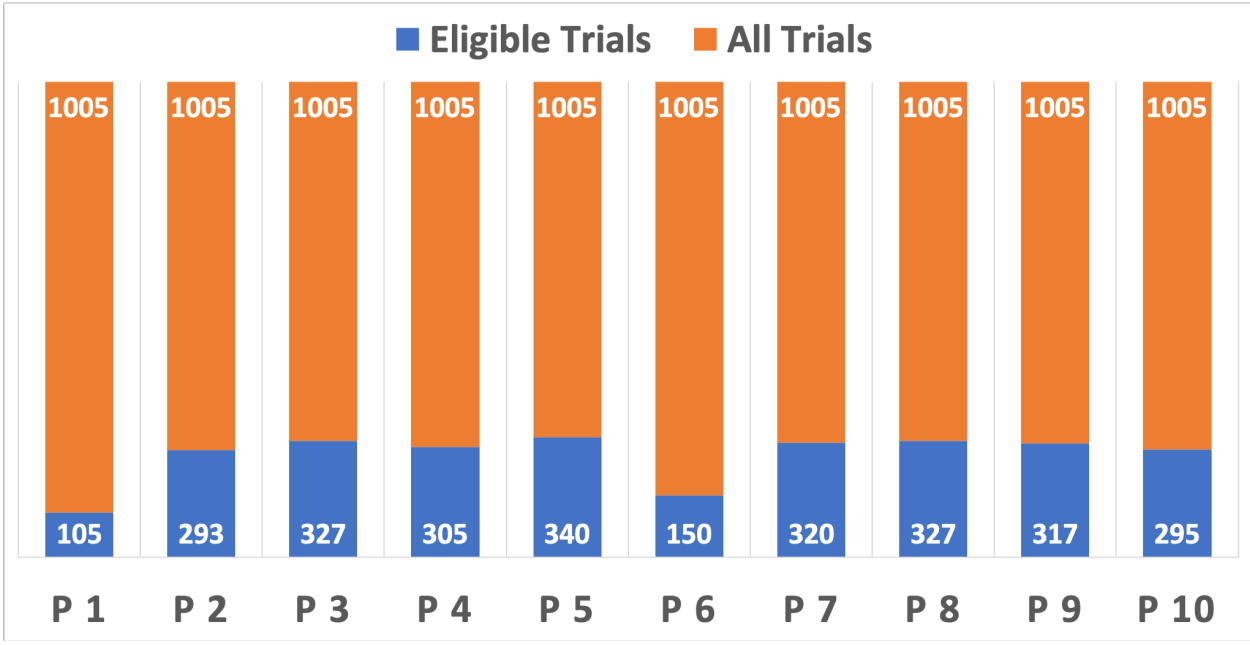


Figure 5.2 Number of eligible trials from the conditional logic algorithm for the first 10 patients

After cleaning these resulting trials for each patient by removing any empty descriptions

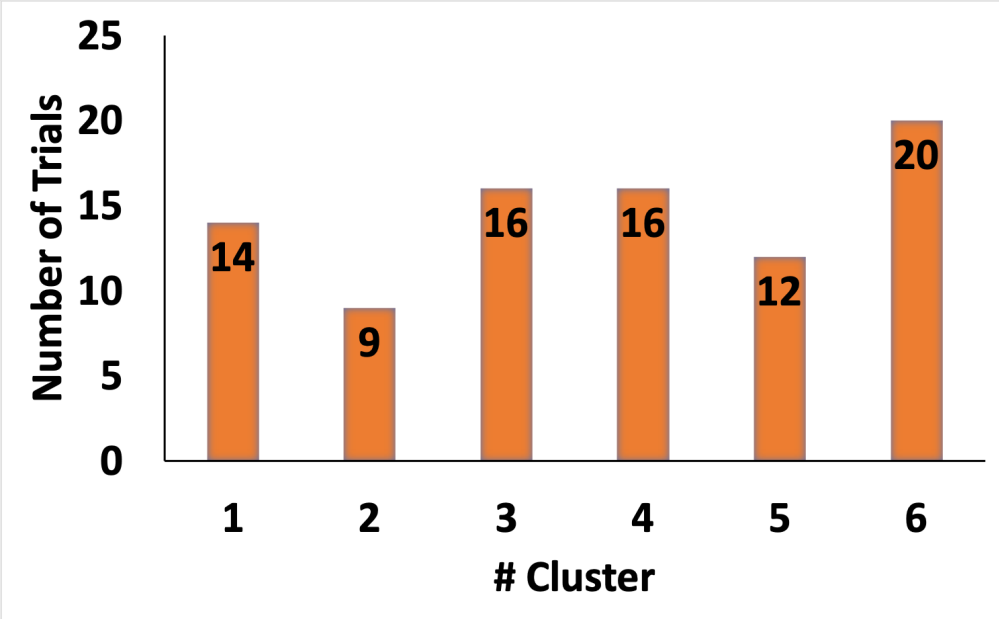


Figure 5.3 Number of Clinical Trials in each 6 Clusters of Patient 1

of clinical trials and using the equation 5.2, we automatically selected the number of clusters for each of the 100 patients. This reduced the number of trials for Patient 1 to 83 trials and 6 clusters. Figure 5.3 shows the number of clusters for Patient 1, dynamically allocated using Equation 5.2, along with how many trials each cluster contains. Once we have clusters for each patient, we took the top five most repeated words in each cluster. Figure 5.4 shows the most common words found in each of the corresponding clusters.

Figure 5.5, left, shows the overall clusters scatter plot for Patient 1 which is the result of agglomerative clustering. Principal Component Analysis (PCA) was used for visualization purposes to illustrate the distribution of each cluster in the first two principal components. Although we are only showing a 2D scatter plot here, there is a distinct separation between the clusters that are shown in Figure 5.5 separated by different colors. Figure 5.5, right, shows the number of times the presented keywords repeated in the selected cluster after taking the three most common keywords in all clusters, e.g. "HIV", "Hemoglobin", and "Platelets" out of the accepted keywords in our algorithm.

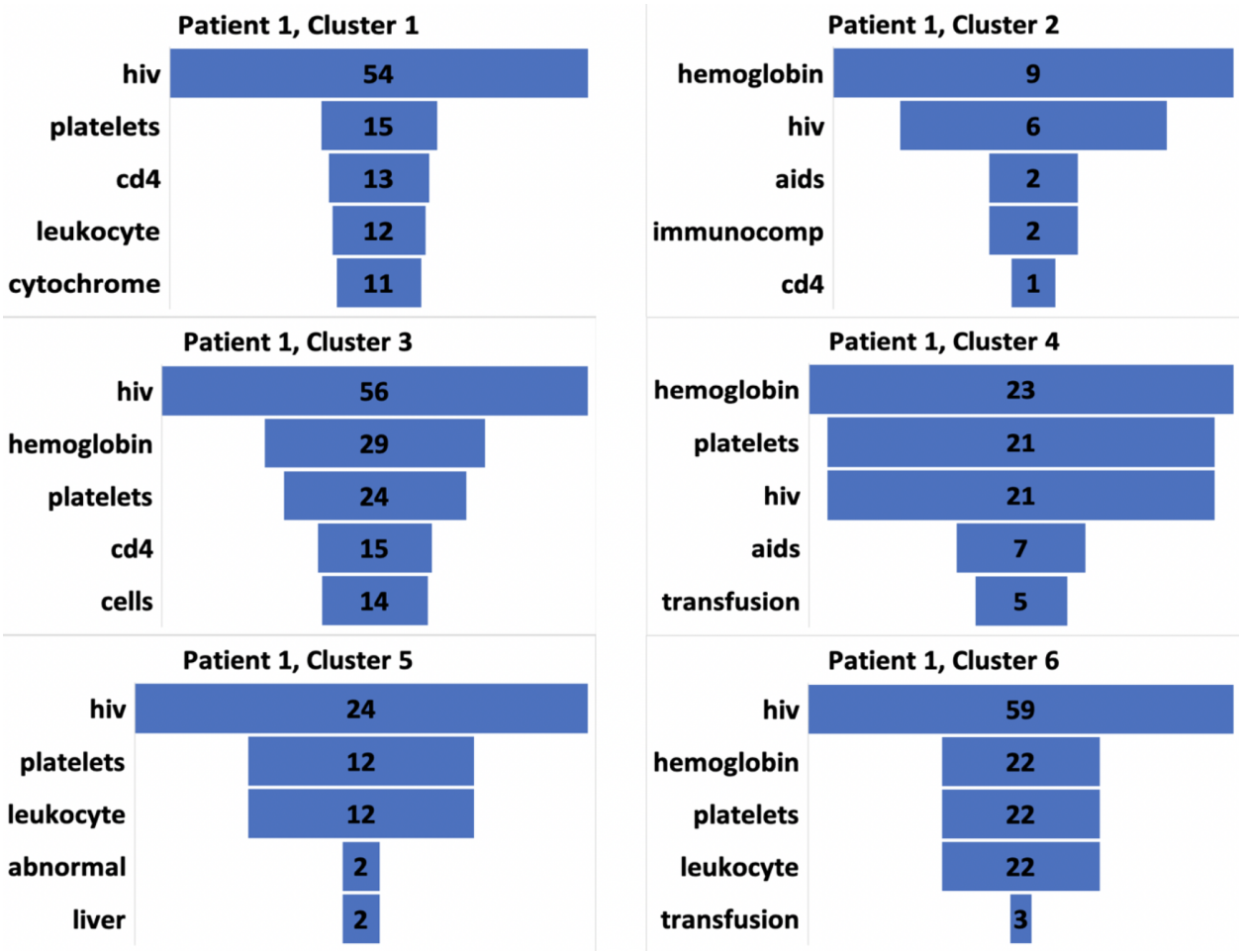


Figure 5.4 Most frequent keywords in all six clusters of Patient 1

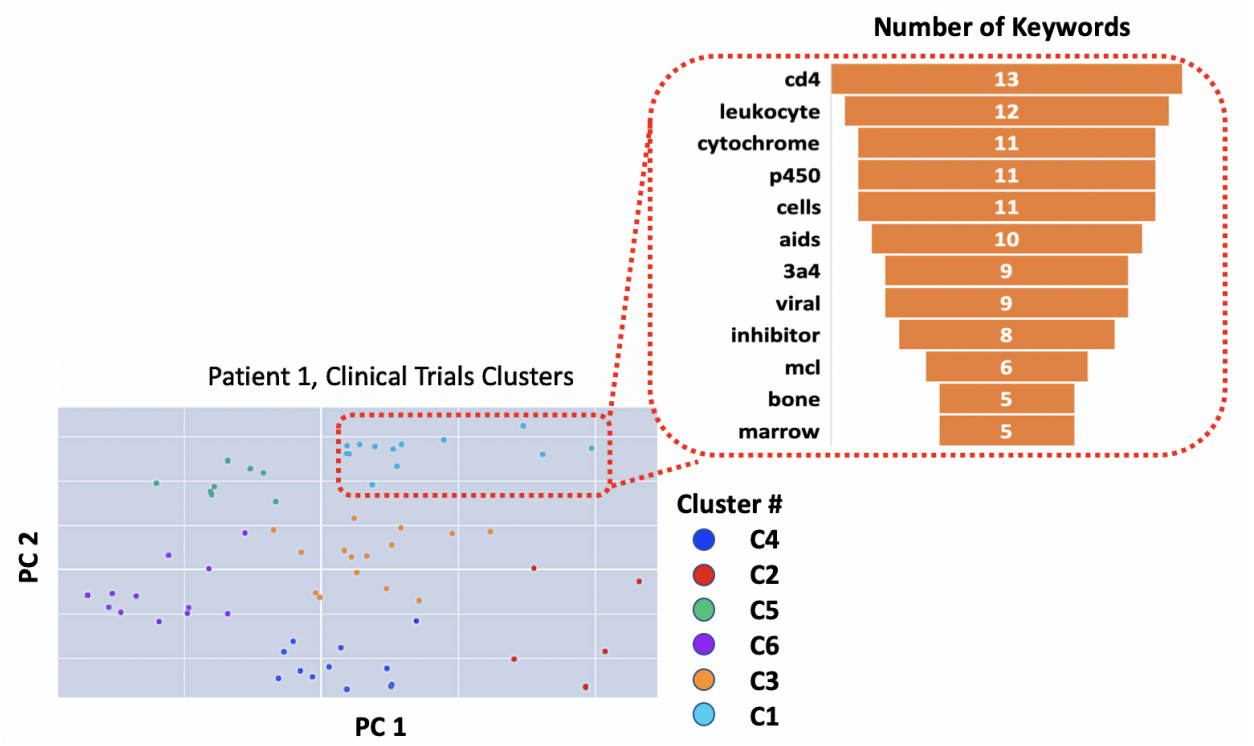


Figure 5.5 Scatter plot of all 6 clusters for Patient 1 and the most common keywords in cluster 1

5.5 Discussion

With the expansion of the number of available clinical trials available for clinicians and other health providers, it is almost impossible to choose the right treatment without spending hours to narrow down the choices. Machine learning techniques have begun to be used for optimizing this process. In our approach, there is a noticeable improvement when comparing the number of clinical trials that doctors have to go through before and after applying our algorithm. The results from the logical comparison presented here significantly narrows down the choices to about third on average for our pool of 100 patients. This was done by simply going through all the eligibility criteria and combine that with individual patient info to select the trials that the patient does not qualify for. This brute force approach alone yields valuable information and can increase efficiency by up to 300%. Alongside this method, the agglomerative clustering, which is a type of an unsupervised learning technique in machine learning, can further facilitate the clinical trial matching process by grouping the similar

text derived from the trial descriptions.

The most frequently used words per cluster are provided to further inform doctors about each cluster so they can visually see the differences as well as use this type of categorization to make their decision and quickly gain insight into the types of trials being returned for the patient. As shown in Figure 5.4, each cluster has common words embedded in them. Looking at the most frequently used words in trials for patient 1, all the words from the trials that are separated in each cluster are strongly related to HIV. For example, cluster 1 contains 54 occurrences of the word "HIV". Also, it has 13 occurrences of the word "cd4" which is also related to HIV. While all the clusters share some common words, there are other unique words that are not found in some clusters. The three keywords of HIV, Hemoglobin, and Platelets were commonly repeated across all the six clusters presented in Figure 5.4 for the patient 1. That is mainly due to the fact that these keywords are parts of the eligibility criteria included in our analysis. Taking those keywords out can help with creating more distinct keywords among clusters, Figure 5.5, right. There is also clearly a need for clinicians to review more than the top 5 keywords presented in Figure 5.4. As also illustrated in Figure 5.5, right, even the least frequent keywords, such as bone, marrow, and 3a4 can be very meaningful features of each cluster. In addition, by increasing the number of patients and the clinical trials, unsupervised learning is able to provide a better categorization of similarities in larger data sets which is required for future precision medicine applications. Ultimately, the HeaLTH algorithm provides a quick and easy approach to patient trial filtration and identification for clinicians and patients alike.

5.6 Conclusion

Utilizing a combination of brute-force logical comparison and machine learning clustering and classifying, our team has created an algorithm that significantly reduces the available trials for a patient-based on personal data matching, and uses hierarchical clustering to

further simplify trial selection and examination. Doctors can use this algorithm to better identify the types of trials a particular patient is more likely to be assigned to, as well as filter out any trials that may or may not yield worthwhile results.

5.6.1 Limitations

The primary limitation of this project was the lack of additional patient data for testing. Furthermore, this algorithm is built around the way that the clinical trials were presented and may prove difficult to implement in a separate environment where clinical trials are presented in a different manner, e.g., if new clinical trials do not have specific inclusion/exclusion criteria presented in a conditional format.

5.6.2 Future Work

Future implementations of this project would be to further streamline the trial selection process for users. This can be accomplished by implementing a user interface with the algorithm that takes in the patient data and directly returns the clustered trials in an easy to read format. Additionally, the clustered patient trials can be directly compared to hand-picked trials for patients selected by clinicians to assist in further refinement and validation of trial selection for patients.

CHAPTER 6
SMART ENERGY IN RESIDENTIAL SECTOR

6.1 Introduction

Excessive ambient temperatures negatively affect the available capacity of most power grid components such as generators, transformers and overhead lines. To make matters worse, this reduction in generation and power transmission capacity often coincides with excess demand on the network, mostly attributed to the over-utilization of air conditioning (A/C) units. Specifically, the A/C demand accounts for approximately 6% of all the electricity produced in the United States, at an annual cost of $29 billion to the residential homeowners [82]. Although in the southern parts of the U.S. such as Texas and its neighboring states, air conditioning accounts for an even greater share of home energy use (18%) compared to the U.S. average (6%) [83]. With the increasing temperatures around the world due to climate change, the air conditioning usage has increased considerably [84], and is one of the factors responsible for higher energy demand on the system, energy fluctuations, and reduction in available power generation reserves [85–87]. This could introduce vulnerability in the power grid and could jeopardize its ability to maintain the balance between generation and demand, which is critical for system stability. Demand response (DR) has been adopted by many electric utilities in emergency situations as an effective tool to counteract the volatility in demand and to compensate for the shortage in generation. The objective in DR is to reduce the demand on the consumer side in exchange for financial incentives. In essence, DR is viewed as a voluntary load shedding mechanism. One way to implement DR is to remotely control and/or shut down some A/C units for a certain period of time to help mitigate the

stress on the grid and to relieve localized congestion [88]. Although DR has shown its effectiveness [89], performing DR under extreme temperature conditions is a delicate matter that requires further analysis. This is because the ability to maintain an acceptable indoor temperature has been the main reason for reduction of heat-induced mortality rates during heat wave events. This is in particular important for households with elderly residents and children. Thus, while DR is an important grid service component that will likely become more important as more renewables become part of the generation portfolio, the effectiveness of DR strategies is contingent upon a proper knowledge of the electric cooling energy use. Naturally, A/C power can be easily measured and monitored; however, privacy reasons often prevent utilities from accessing this data, which would only be accessible behind the meter. This is why it is desired to estimate the A/C demand in a non-intrusive fashion without violating the privacy of the residents. This information can be very beneficial for the electric utility when trying to manage demand in congested regions of the power grid. Finding a solution for estimating residential A/C demand is the goal of this paper. A novel Non-Intrusive Load Monitoring (NILM) technique is proposed here to decompose the measured total power consumption of a house into A/C versus non-A/C power consumption. The algorithm proposed in this study allows for detecting the activity cycle of an A/C unit and estimating its energy consumption. This information can then be used by the electric utility to implement (or customize) a DR event, while ensuring the privacy of the residents is not violated.

6.2 Related Work

Several load disaggregation techniques have been reported in the literature. For example, in [90], power consumption of each appliance is modeled by a Hidden Markov Model (HMM) while the aggregated demand is modeled by factorial HMM. Other investigations [91,92] also utilized factorial HMM to discover the ON/OFF state of the appliances as a solution for a

non-intrusive approach. However, due to several disadvantages of HMM, the approach often fails to represent appliances with a continuous fluctuating power demand [92], making them less effective for capturing the accurate behavior of A/C units.

Time series analysis techniques are another class of solutions that can be used for this purpose. Here, the idea is to decompose the data into various components and then use this information for prediction purposes. Forecasting techniques vary from the simple averaging solutions and exponential smoothing methods to more complicated Auto Regressive Integrated Moving Average (ARIMA) models. In particular, ARIMA models have been widely used for load forecasting with significant results [93]. Although ARIMA models have been shown to be effective in time series analysis, they have their own disadvantages. In particular, often times the structure of the ARIMA model and the orders of the moving average (MA) part and the auto-regressive (AR) part are chosen either subjectively based on the experience of the programmer/forecaster or using trial and error. In addition, ARIMA models are known to be "backward looking" when it comes to predicting future values. Because of this reason, they are considered to be generally poor performers at predicting turnings [94]. This could be a major limitation in forecasting A/C consumption due to its intermittent ON/OFF cycles. Furthermore, ARIMA offers a linear model, which may not be sufficient for modeling some complicated nonlinear patterns in the data.

To overcome the main problems introduced by HMM and ARIMA, many of the recent studies have been shifting their focus to deep learning. Deep learning is the most popular topic in the field of computer science, and has demonstrated and produced significant achievements in computer vision, natural language processing, and forecasting, to name a few. The basis of deep learning is to use multiple processing layers with convoluted structures and non-linear operations to extract high-level complex abstractions [95, 96]. By exploring these techniques, we are able to extract detailed information about the power consumption of residential buildings as well as the underlying A/C usage patterns. A recent study done by [97] proposed a solution for managing national level DR by introducing Long-Short

Term Memory (LSTM) recurrent neural networks to produce load forecast at the national level. Another study by [98] also proposed a LSTM based approach to forecast monthly electric demand of the residential sector. Hybrid models have also been proposed to forecast electricity consumption in residential and commercial buildings. For instance, the authors of [99,100] proposed a hybrid model which consisted of convolutional neural networks (CNN) and LSTM. Another approach proposed by [101] utilized a hybrid model which consisted of CNN combined with LSTM and autoencoder (AE) to forecast electricity consumption in residential and commercial buildings. While these studies provided useful insights into energy consumption patterns and forecasting in residential buildings as a whole, analyzing A/C demand patterns in these buildings is relatively less explored. Nevertheless, some researchers have looked into the problem of forecasting A/C load in residential buildings. A study done by [102] proposed Levenberg-Marquardt algorithm based artificial neural networks to carry out short term A/C load forecasting. Similarly, the authors of [103] employed a support vector machine (SVM) based model to achieve the same. The summary of highlighted studies are shown in Table 6.1.

Table 6.1 Highlighted studies for load forecasting and disaggregation

Paper	Category	Proposed Algorithm
[90–92]	Statistical	HMM
[93]	Statistical	ARIMA
[99–101]	Hybrid	CNN-LSTM, LSTM-AE
[97, 98, 104]	DL-based	LSTM

In our previous study [104], we used LSTM to classify A/C usage patterns and forecast future A/C loads purely based on the total power consumption. Hence, the significance of temperature set-point was not considered. In this study, the impact of controlling the A/C temperature set-point of residential buildings is carefully examined in order to execute DR properly. The following are the main contributions of this research work:

- Use of Long-Short Term Memory (LSTM) to forecast A/C electric energy use based on one-minute interval smart meter data.
- Develop and calibrate three EnergyPlus building energy models using actual building energy data to fully understand how energy is used in the analyzed homes.
- NILM validation using data from calibrated energy models
- Examine the impact of controlling the A/C unit on the temperature inside the buildings.

6.2.1 Data and building physics based energy modeling

We use house data provided by Pecan Street Inc. to develop, simulate, and calibrate building energy models. These data include one-minute smart meter measurements for: total house power consumption, A/C power consumption, and power consumption for different appliances (i.e. dryer, stove, microwave, oven, and etc.) for certain houses in the Mueller district, Austin, TX. Specifically, we selected three buildings due to the relatively complete data set. These buildings have an identification number of 2470, 2814, and 3367 in the Pecan Street Inc. database and are referred to with the same identification number throughout this publication. Buildings 2470, 2814, and 3367 were built in 2008, 2009, and 2007, respectively and are fairly new buildings. Pecan Street Inc. house data also included the house energy audit data containing: number of floors, bedrooms, various measurements of the house geometry such as number, area, and orientation of windows and doors, and infiltration value (ACH). However, it does not include thermostat setpoints, envelope assemblies or internal mass. Thus, we developed detailed building energy models to fully understand the different parameters in each analyzed building. An example of Pecan Street house audit data is shown in Table 6.2.

Data preprocessing was applied to modify the raw data from the Pecan Street Inc. into a suitable form for both classification and regression purposes. We utilized Python NumPy

Table 6.2 Example of Pecan street house audit data

house id	number of levels	number of bedrooms	conditioned area (ft^2)	ACH50	...	type of home
2470	2	3	1544	7.5	...	single family
2814	2	3	1842	3.2	...	single family

array functions (e.g., reshape, flatten, and squeeze) to transform the dimension of the data into an appropriate input shape for the proposed deep learning model.

Robust NIML model validation requires multiple buildings to increase confidence in the proposed method. Thus, this study developed and calibrated building energy models for all houses with available audit data, appliances, whole house, and A/C power consumption data and later modified to increase the number of sample data sets (and diversity of buildings) from three to six buildings. Buildings are simulated using 2014 actual meteorological year (AMY) data and National Renewable Energy Laboratory (NREL)'s Building Energy Optimization (BEopt) tool [105]. BEopt is a simplified graphical user interface (GUI) for EnergyPlus, the U.S. Department of Energy (DOE)'s whole building physics based energy modeling program [106]. BEopt allows fast construction of 3D model of the building but has some limitations compare to EnergyPlus. Thus, model development and preliminary calibrations are done in BEopt and detailed calibrations are performed with EnergyPlus.

Model calibration is required since not all input parameters such as thermal mass and thermostat setpoints are known. This study varies the following unknown parameters to calibrate the models against the measurement data from Pecan street Inc: thermostat setpoints, zone thermal capacitance multiplier, temperature difference between cutout and setpoint, wall insulation, and window properties (U-value and SHGC). For thermostat setpoints, we use data from previous studies to limit the range of potential setpoints [107] and [108]. Every building has a different characteristic in terms of thermal mass which consequently affects the internal zone temperature fluctuations. Internal thermal mass of the modeled buildings is simulated in EnergyPlus using two approaches: a) internal zone capacitance multiplier that provides a calibration input as it multiples the mass provided by interior air and b)

internal mass object that integrates mass of elements (i.e. furniture, walls, books) and their associated surface areas. BEopt assumes a value of one for internal zone capacitance of every modeled building; however, considering the on/off characteristics of the HVAC system, the internal thermal mass object is modified for every building [109] to accurately capture the temperature fluctuations of each thermal zone and consider thermal mass.

EnergyPlus model validation follows ASHRAE Guideline 14 metrics including Coefficient of Variation of Root Mean Square Error (CV-RMSE) and Normalized Mean Bias Error (NMBE) for total and A/C power consumption [110]. ASHRAE Guideline 14 suggests that for hourly calibrations, CV-RMSE and NMBE should be less than 30% and 10%, respectively.

6.2.2 Simulated buildings

Once the simulated buildings are calibrated, any changes within the building (e.g. wall insulation, infiltration, etc.) creates a different simulated building as it has a different thermal performance than the original building. This is done to represent a diversity factor between the buildings built in a geographically cohesive neighbourhood (i.e. Austin's Muller district in this study). This diversity factor includes a broad range of variations in building internal loads (due to variations in occupants and equipment) and building characteristics such as window properties (SHGC and U-value), wall insulation, and infiltration. These variations are done considering allowable ranges provided by International Energy Conservation Code (IECC) [111]. Table 6.3 provides parameters changed in each real building model to develop the modified ones; hereafter, referred to as 2470-M, 2814-M, and 3367-M. These new simulated buildings' total and A/C electric consumption results provide data for six simulated houses (3 real and 3 modified) for June 1st, 2014 to August 31st, 2014 (summer season) which is then used as the input for the non-intrusive A/C load disaggregation method.

Table 6.3 Multipliers used for variables modification in virtual buildings compared to real ones

Variable	2470-M	2814-M	3367-M
Internal loads	-	-	1.2
Wall insulation	-	1.6	-
Infiltration	0.5	-	-

6.2.3 Algorithms

6.2.3.1 Baseline Methods

Support Vector Machine (SVM): SVM is a supervised machine learning algorithm which can be used for both classification and regression purposes. SVM has been studied extensively and applied to various problems such as pattern classification and function approximation [112]. In SVM, each point is rearranged in n-dimensional space and SVM tries to find the hyperplane that separates the two classes [113]. Thus, SVM is an example of a linear two-class classifier, where the labels are +1 (positive samples) or -1 (negative samples) [114].
Random Forest (RF): Random forest is a popular machine learning algorithm that is used to solve both classification and regression problems. Random forest is an extended version of bagging method which provides an additional layer of randomness to bagging. In standard trees, each node is split using the best split among all variables. However, in a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. Random forest is also easy to use in the sense that only two parameters are required and is usually not sensitive to their values [115, 116]. In order to develop a random forest model, the following parameters have been considered:

- N_estimator: the number of trees in the forest.
- Criterion: A function that measures the quality of a split in the random forest.
- Max depth: The longest path between the root node and the leaf node.
- Minimum sample split: the minimum number of sample required to split an internal node in the tree.

The parameters used for random forest are shown in Table 6.4.

Table 6.4 Parameters for the baseline random forest model

N_estimators	Criterion	Max Depth	Minimum Sample Split
100	Gini Impurity	None	2

Auto Regressive Integrated Moving Average (ARIMA): ARIMA models are the most general method for forecasting a time series data that can be made to be stationary by differencing with a nonlinear transformation methods such as logging. A nonseasonal ARIMA can be classified as an "ARIMA(p,d,q)" model, where p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and q is the number of lagged forecast errors in the prediction equation [117]. Table 6.5 shows the parameters for the ARIMA method.

Table 6.5 Parameters for the baseline ARIMA model

Parameters	Input
p	2
q	1
d	0

6.2.3.2 Proposed Method

Long-Short Term Memory (LSTM)Recurrent neural network has gained enormous attention due to its ability to handle time series data. Recurrent networks utilize their feedback connections to store recent input information. However, the conventional recurrent neural work that uses backpropagation through time [118] could lead the gradients to blow up or

vanish [119, 120]. LSTM is a special type of recurrent neural network that was designed to prevent gradients from exploding or vanishing by storing information for long periods of time [119, 120]. The LSTM utilizes memory block that contains one or more memory cells. The memory cells have self-loops that allows them to store temporal information encoded on the cell's state, which is important for numerous sequential tasks. A general LSTM memory block architecture is illustrated in figure 6.1 [121]. The state of cell is split into two vectors, $h_{(t)}$ and $c_{(t)}$, where $h_{(t)}$ can be explained as the short-term state and $c_{(t)}$ as the long-term state. The current data x_t and hidden short-term state from previous timestep h_t are processed by different gates that does different operations, which allows LSTM to be capable of adding or removing information to the cells state. These operations are carried out by three gates: the input gate, the forget gate, and the output gate [97, 119, 120, 122]. The input gate controls the amount of input data flowing into the cell and the decision of adding parts of $c_{(t)}$ to the long-term state. The input gate is controlled by $i_{(t)}$. The forget gate controls the amount of data that should be erased. The forget gate is controlled by $f_{(t)}$ Based on these information, the cells output is generated by the output gate to control the extend to which the value in the cell is used to compute the output activation of the LSTM unit. The output gate is controlled by $o_{(t)}$ [119]. The following equations summarize how each gate operates according to its functionality:

$$i_{(t)} = \alpha(W_{xi} * x_t + W_{hi} * h_{t-1} + b_i) \tag{6.1}$$

$$f_{(t)} = \alpha(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \tag{6.2}$$

$$o_{(t)} = \alpha(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \tag{6.3}$$

$$c_{(t)} = tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \tag{6.4}$$

$$c_{(t)} = f_{(t)} \times c_{t-1} + i_{(t)} \times c_{(t)} \tag{6.5}$$

$$h_{(t)} = o_{(t)} \times tanh(c_{(t)}) \tag{6.6}$$

Where:

- W_{xi}, W_{xf}, W_{xo}, and W_{xc} are the weights for the input gate, forget gate, output gate and the memory cell state.
- b_i, b_f, b_o, b_c are the bias terms for the input, forget, output gate and the memory cell state.

We implemented the LSTM model using Keras, a deep learning library written in Python [123]. NVIDIA GPU was utilized to maximize computational process. Also, we have selected mean squared error (MSE) as the loss function, and Adam optimization was utilized. A dropout layer was considered in the LSTM model in order to prevent over-fitting. The details of the architecture used by our LSTM model are shown in table 6.6.

All computations were performed on a desktop PC with Intel Core i7-4790 CPU (4x3.60GHz), 16GB DDR4 RAM, and NVIDIA GeForce GTX 1060 6GB GPU.

Table 6.6 Proposed LSTM Deep Learning Architecture

Layer	Type	Node	Activation
1	LSTM	120	Tanh
2	LSTM	60	Tanh
3	Dense	1	-

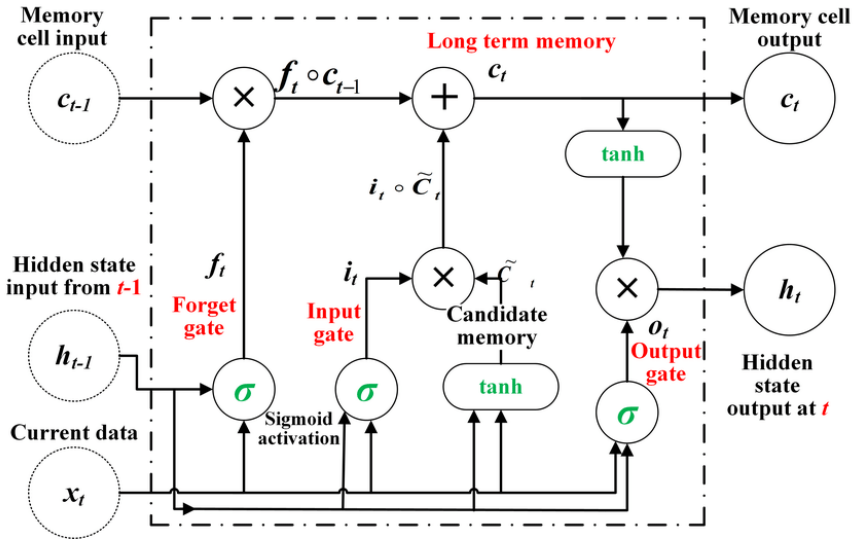


Figure 6.1 Illustration of a LSTM unit

6.3 Results

6.3.1 Building physics based energy modeling

Table 6.7 provides a summary of items modified either in BEopt or EnergyPlus to calibrate the total and A/C power of the buildings 2470, 2814, and 3367. It also shows final values used to achieve the highest accuracy which are within the ranges of expected values.

Table 6.7 Variables used for buildings' EnergyPlus model calibration

Variable	2470	2814	3367
Cooling setpoint (F)	76 (9 AM-13 PM) and 70 (other)	74	78 (7-10 AM) and 74 (other)
Heating setpoint (F)	68	68	64
Zone thermal capacitance multiplier (-)	2.0	7.0	2.0
ΔT between cutout and setpoint (C)	0.95	1.0	0.9
Window U-value ($\frac{Btu}{hr-ft^2-F}$)	0.65	0.3	0.3
Window SHGC (-)	0.21	0.29	0.11

Figures 6.2 and 6.3 show results from EnergyPlus aginst measured data for buildings 2814 and 2470's A/C power for June 21st, 2014. For this particular day of building 2814, Figure 6.2 shows that the simulated results are slightly overestimated from midnight till 10:00 AM and from 5:00 - 11:59 PM. Conversely, the model results are slightly underestimated

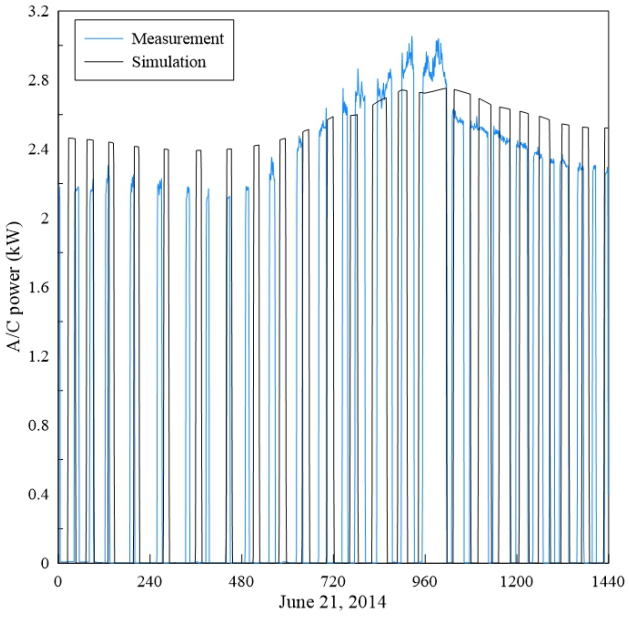


Figure 6.2 Data comparison for building 2814

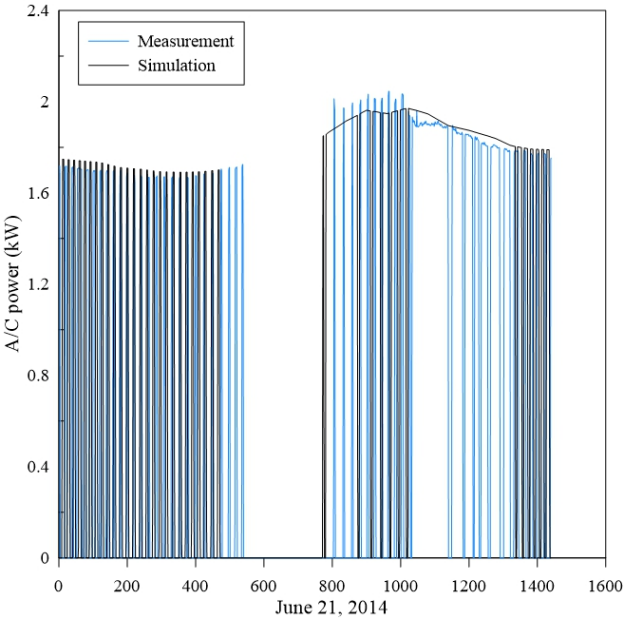


Figure 6.3 Data comparison for building 2470

from 10:00 AM till 5:00 PM. However, EnergyPlus is able to accurately predict the ON/OFF status of the HVAC system due to the even distribution of the cooling setpoint throughout the day (see Table 6.7). For the same day of interest, building 2470 shows different results (see Figure 6.3). Although the magnitude of the measurement and simulated data are the same and the shape of the simulated data follows measurement, but due to the complex cooling setpoint (see Table 6.7), sometimes there are a few minute delay between actual ON/OFF change of status and simulated ON/OFF for each day. Figures 6.2 and 6.3 only show results for one day (June 21, 2014) but Table 6.8 presents CV-RMSE, NMBE, and RMSE for the modeled buildings for the entire analyze period. All simulated houses meet ASHRAE Guideline 14 requirements, although the models are carried in a minute level in this study.

As shown in Table 6.3, multipliers applied to various diversity factors (internal loads, wall insulation, and infiltration) in each building to increase the number of building data sets available. Any changes in the building model resulted from the diversity factors changes the associated total and A/C consumption behaviour. For instance, Figure 6.4 presents the

Table 6.8 CV-RMSE (%), NMBE (%), and RMSE for buildings 2470, 2814, and 3367

Building	2470	2814	3367
CV-RMSE	0.48	0.37	0.34
RMSE	0.55	1.6	0.67
NMBE	10	5.4	7.9

A/C consumption behavior comparison for buildings 2814 and 2814-M. As discussed in Table 6.3, building 2814-M has a different wall insulation (with a multiplier of 1.6) compared to building 2814. Accordingly, the A/C power consumption magnitude is decreased for the modified building and there is a slight shift in the peak time considering the increased insulation.

6.3.2 A/C Activity Cycle Determination and Estimation

For the purpose of performance evaluation, we have selected the data from June 1st, 2014 to July 31st, 2014 as the training set (86,400 data points) and the remaining data (August 1st, 2014 to August 31st, 2014) was used as the testing set.

The following metrics were used for algorithm evaluation:

$$TP = number\ of\ true\ positive \tag{6.7}$$

$$FP = number\ of\ false\ positive \tag{6.8}$$

$$FN = number\ of\ false\ negative \tag{6.9}$$

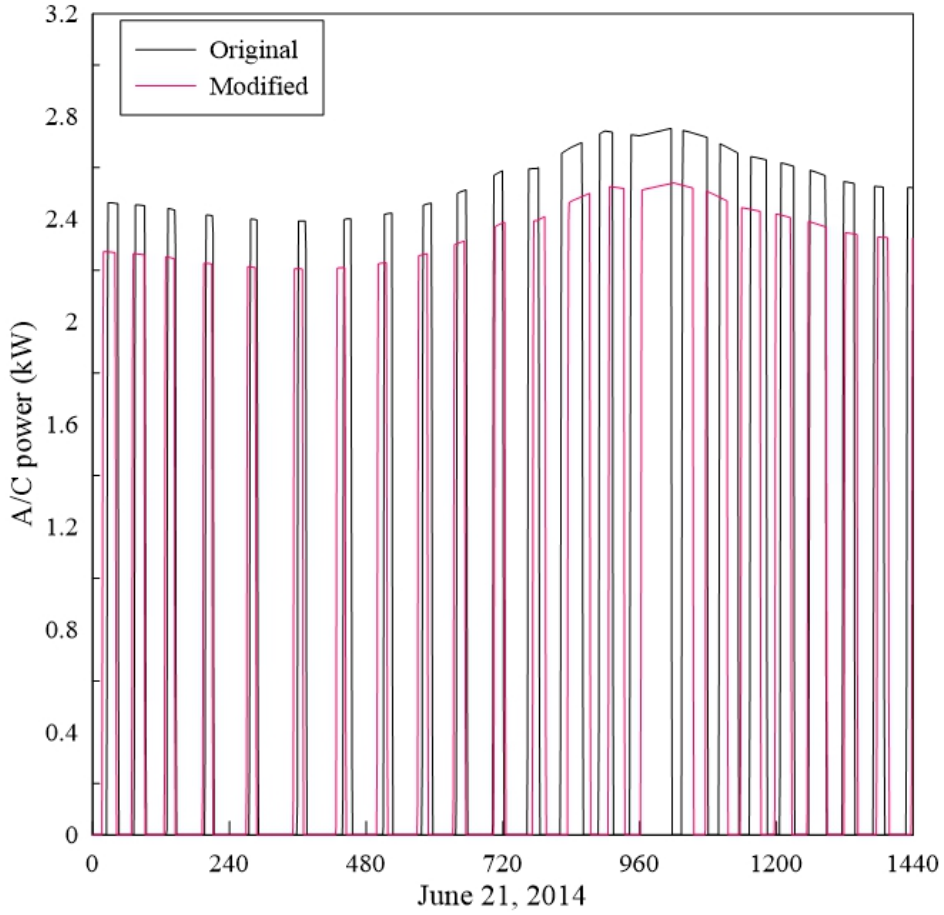


Figure 6.4 Comparison of A/C power for original and modified EnergyPlus models of building 2814

$$P = number\ of\ positives\ in\ ground\ truth \tag{6.10}$$

$$N = number\ of\ negatives\ in\ ground\ truth \tag{6.11}$$

$$y_t = actual\ power\ consumption\ at\ time\ t \tag{6.12}$$

$$\bar{y}_t = estimated\ power\ consumption\ at\ time\ t \tag{6.13}$$

$$Accuracy = \frac{TP + TN}{P + N} \tag{6.14}$$

$$Precision = \frac{TP}{TP + FP} \tag{6.15}$$

$$Recall = \frac{TP}{TP + FN} \tag{6.16}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{6.17}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_t)^2} \tag{6.18}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - \bar{y}_t| \tag{6.19}$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \tag{6.20}$$

Equations (8) $\sim$ (11) were used to evaluate the classification performance, and equations (12) $\sim$ (14) were used to evaluate the estimation performance. table 6.9 shows the performance for the A/C ON/OFF activity determination. The accuracy of A/C ON/OFF activity determination for the baseline methods (i.e., SVM, RF) and the proposed algorithm (i.e., LSTM) achieved over 95% for all six houses, with the average accuracy of 0.972 (SVM), 0.961 (RF), and 0.984 (LSTM). The minimum/maximum accuracy for the houses were 0.969/0.976 (SVM), 0.951/0.967 (RF), and 0.981/0.986 (LSTM). LSTM achieved better

Table 6.9 Performance of A/C ON/OFF activity classification

SVM House	Accuracy	Precision	Recall	F1 Score
1	0.976	0.977	0.982	0.971
2	0.974	0.982	0.974	0.978
3	0.969	0.957	0.987	0.972
4	0.972	0.960	0.986	0.973
5	0.971	0.947	0.993	0.970
6	0.969	0.944	0.991	0.967
Average	0.972	0.961	0.986	0.972
RF House	Accuracy	Precision	Recall	F1 Score
1	0.967	0.977	0.966	0.971
2	0.969	0.979	0.968	0.973
3	0.953	0.966	0.946	0.956
4	0.951	0.964	0.936	0.950
5	0.963	0.982	0.938	0.959
6	0.965	0.984	0.937	0.960
Average	0.961	0.975	0.949	0.962
LSTM House	Accuracy	Precision	Recall	F1 Score
1	0.982	0.979	0.99	0.985
2	0.985	0.985	0.99	0.988
3	0.982	0.978	0.988	0.983
4	0.986	0.981	0.991	0.986
5	0.986	0.985	0.984	0.985
6	0.981	0.973	0.986	0.979
Average	**0.984**	**0.980**	**0.988**	**0.984**

performance than the baseline methods in terms of accuracy, precision, recall, and f1 score. Figure 6.5 shows examples of classified A/C ON/OFF activity classification by LSTM.

The performance of the A/C power consumption estimation by ARIMA and LSTM are shown in table 6.10. The R^2 value of A/C power consumption estimation for all six houses were above 0.85, with the mean R^2 value of 0.903 (ARIMA) and 0.910 (LSTM). The minimum/maximum R^2 values for all houses were 0.876/0.950 for ARIMA and 0.893/0.929 for LSTM. When comparing the performances of ARIMA and LSTM in terms of RMSE, LSTM had slightly less error than ARIMA. Our proposed LSTM model outperformed the baseline methods in classification problem, and achieve slightly better performance in regression problem.

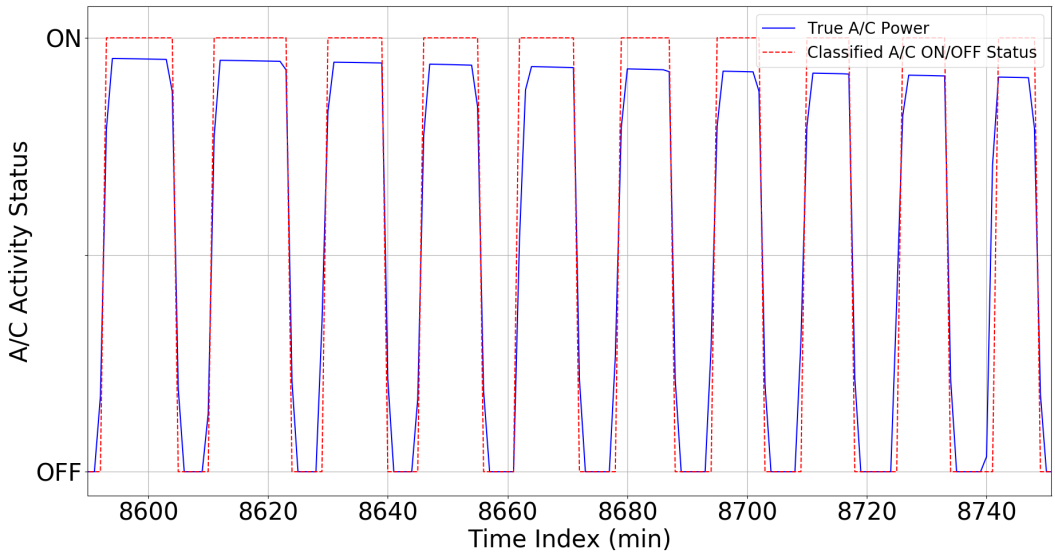
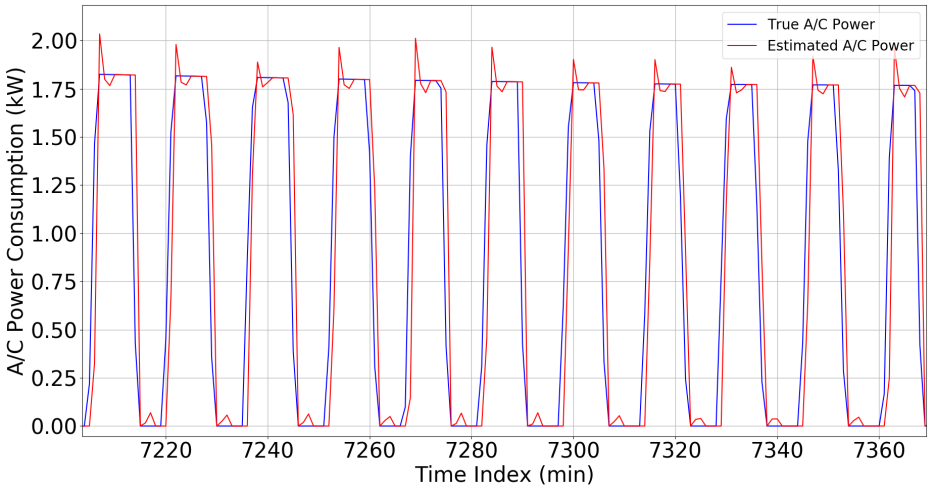


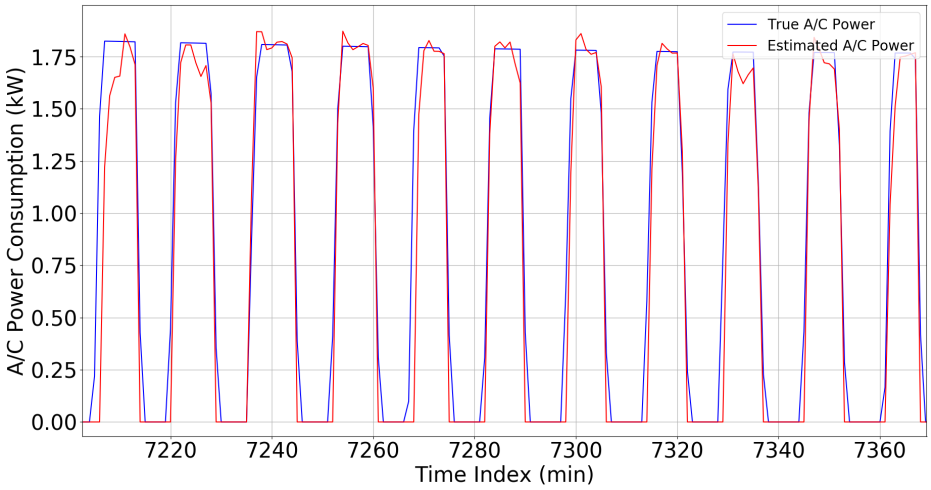
Figure 6.5 Classified A/C ON/OFF activity status by LSTM

Table 6.10 Performances of A/C Consumption Estimation

ARIMA House	RMSE	MAE	R^2
1	0.30	0.10	0.890
2	0.360	0.123	0.880
3	0.537	0.180	0.880
4	0.548	0.190	0.876
5	0.304	0.067	0.950
6	0.343	0.080	0.940
Average	0.399	0.123	0.903
LSTM House	RMSE	MAE	R^2
1	0.310	0.161	0.893
2	0.30	0.164	0.923
3	0.499	0.282	0.896
4	0.486	0.282	0.902
5	0.381	0.165	0.919
6	0.384	0.183	0.929
Average	**0.393**	0.206	**0.910**



(a)



(b)

Figure 6.6 Estimated A/C power consumption by ARIMA (a) and LSTM (b)

Figure 6.6a shows an example of estimated A/C power consumption by ARIMA and figure 6.6b shows an example of estimated A/C power consumption by LSTM.

For situations with limited known building parameters, deep learning method (e.g. LSTM) captures the ON/OFF state of the A/C system more accurately compared to EnergyPlus because EnergyPlus needs more than 20 different building parameters to obtain accurate results. However, the proposed method only require the total energy use and the A/C energy for a model training purpose and they only uses the total energy use to make estimations.

If all building parameters are known, EnergyPlus is more robust option with results valid for longer period of times (year) than ARIMA and LSTM (weeks). EnergyPlus performance will remain almost constant even when significant disturbances (interior or exterior) are introduced which is not the case for deep learning methods or any other reduced order model. However, EnergyPlus's accuracy comes with additional cost in terms of the resources required for model development.

6.4 Conclusion

The last decades have witnessed relentless changes in weather patterns. Abundant growth of power consumption mostly for cooling purposes and failure to meet the energy needs due to fuel shortage and capacity limitations have been shown to lead to failure in operation and availability of critical infrastructures. Demand Response (DR), especially targeting residential A/C units, has been used by many power utilities during critical situations to alleviate the stress on the system. In order to properly execute DR, the cooling load of buildings must be identified. However, accessing the A/C power data can be challenging due to privacy concerns. A non-intrusive approach can be useful in cases like this to breakdown aggregated total power consumption of a residential building into appliance level data, specifically A/C power consumption. This section investigates the effectiveness of deep learning algorithms in determining the A/C activity cycle and estimating the future A/C power consumption of six simulated homes. The proposed deep learning algorithm has shown its capability through high accuracy and consistency in identifying A/C ON/OFF cycles and also estimating its actual energy use compared to physics-based building energy models developed by EnergyPlus. The implementation of our proposed solution can help improve the effectiveness of relevant residential DR programs. At the high level, this can indirectly help with ensuring power system resiliency, reliability, and availability during periods of extreme weather conditions such as heat waves and assuring the supply of power to crucial loads.

CHAPTER 7
SMART ENERGY IN INDUSTRIAL SECTOR

7.1 Introduction

The Tennessee Valley Authority (TVA) is the nation's largest public power provider serving 9 million people in parts of seven southeastern states, through directly served industrial customers and local power companies. With more than 4,000 power billing meters that TVA must account for, management of meter processes is critical in accounting for load transfers, damage, and outages to minimize error and ensure accurate billing. TVA employs teams of analysts whose job is to review daily meter occurrences and report all incidents that deviate from relatively normal operation, which we'll refer to as 'anomalies' that can be attributed to outages, damage, theft, load transfers, etc. Residential energy loads are normally auto-estimated by the meter-reading systems used in the utility industry, but TVA's meters also account for loads from industrial and commercial customers, which are so large that auto-estimation is not feasible. Errors made in the validation process can be costly to the providers or to customers. By use of deep learning algorithms developed in Python, we are able to learn the relative behavior of data in a given time of year with regards to typical energy use patterns. Using the information accumulated from both statistics and deep learning, we can monitor abnormal patterns within energy consumption data.

7.2 Related Work

In recent years, numerous approaches have been made to more accurately attempt energy load forecasting through the use of predictive modeling. Because energy use patterns are dependent on a wide variety of factors, determining an appropriate forecasting model for energy consumption behaviors is a highly specialized task. Therefore, models that are based on network specifications rather than generalization are preferred including: long-short-term memory (LSTM), the autoregressive integrated moving average (ARIMA), and other complementary models, such as vector autoregression (VAR), Bayesian vector autoregression (BVAR), and seasonal ARIMA (SARIMA) [124].

Although ARIMA and VAR models are proficient in forecasting daily load based upon the linear aspect from data, they are unable to account for the nonlinear aspects of the load time series which represents randomness induced by unaccounted emergencies and weather conditions [125]. To uncover the nonlinear aspects of time series, LSTM was used. LSTM is a popular technique of deep learning based on the recurrent neural network (RNN) framework [126]. This new evaluation included comparing a variation of linear and non-linear forecasting models with a hybrid model which uses both linear and non-linear techniques, implemented via the R programming language. It determined that the LSTM RNN architecture outperformed all other models in terms of accuracy when using the metric of mean absolute percent error (MAPE) [125].

Further, it was found that LSTM performed far better than standard empirical and machine learning approaches, such as empirical means, conventional back propagation neural networks, and k-nearest neighbor regressions [126]. Previous research on the task of forecasting energy demand concluded that LSTM was the superior model as opposed to using the ARIMA model and VAR model [125].

Another study performed forecasting on household energy consumption using the BVAR model; BVAR is a variant of VAR with the use of Bayesian methods to estimate vector

regression. Comparisons were made between the accuracy of ARIMA and VAR, with both models producing appropriate results showing the sustained growth of household energy consumption in China [127].

7.3 Contribution

In this paper, we will investigate the combination of LSTM and ARIMA as an effective model for forecasting energy demand. We propose this combination method with the intent of benefiting from the two model's separate abilities, including handling randomness and non-linear parameters as well as utilizing a machine learning approach to handle a larger variety of specialized factors.

Through our work, we believe the combinational use of ARIMA and LSTM will produce more accurate results in anomaly detection of energy consumption. With the aptitude to forecast time series data of ARIMA and the LSTM RNN architecture, our goal with this research is to create a tool automates the reviewal process for energy consumption data, minimizing manual work on a day-to-day basis.

7.4 Methods

7.4.1 Data Processing

This research is focused on energy consumption data from roughly 30 meters providing power to business and residential areas. The anomalies that were found in this region were predominantly from load transfers between meters, done so while one is under maintenance or to help carry a load of power, which resulted in little variation of anomalies detected. Through the use of structured query language (SQL), TVA provided data from Itron Enterprise Edition (IEE) and Oracle Utilities/Lodestar from 2014-2016, measured in kilowatt hours delivered (kWh del), kilovolt-ampere-reactive hours (kVARh) delivered, kVARh re-

ceived, and number of pulses (V2h) for each service point in a given time interval. IEE provides meter management data, which give raw meter readings fed to the algorithm utilizing ARIMA and LSTM models to learn the meters' behavior, whereas Oracle Utilities/Lodestar holds the corrected meter readings. Temperature data for the corresponding dates of data given were also employed in the machine-learning algorithm.

We implemented LSTM using Keras, a deep learning library provided in Python [128]. Through Keras, we also deployed the use of Adam, an algorithm for stochastic optimization [129]. GPU was utilized in order to speed up the deep learning training process. Computations for both ARIMA and LSTM were performed on a desktop PC with Intel Core i7-4790 CPU (4x3.60GHz), 16GB DDR4 RAM, and GeForce GTX 1060 6GB GPU.

Before analysis was performed, data cleaning and preprocessing were required to manage gaps of missing values due to retired meters and new meters. An understanding of meter point ID's and correlating service point ID's was also needed to map IEE and Oracle/Lodestar data together, as one was used in the power sector and the other in billing. The mapping is crucial in the relationships between service points and meter points, through a line of relationships that starts and continues from the service point (physical meter), account, recorder, ending at the meter point, whose name is not always identical to the service point.

A graphical representation of this service point to meter point relationship as it relates to the mapping of IEE and LodeStar data is shown in Figure 7.1.

The conversion of V2h data was necessary because voltage can loosely be used to verify load transfers from outages. When a load transfer occurs, the kWh reading of one meter would show a spike or drop in energy levels delivered while the voltage reading would remain constant simultaneously. Unfortunately, because there are many cases that can use the voltage reading of meters for a form of verification, the use of voltage proved to be imprecise in concluding whether or not an incident occurred, making it an unreliable factor in determining anomalies.

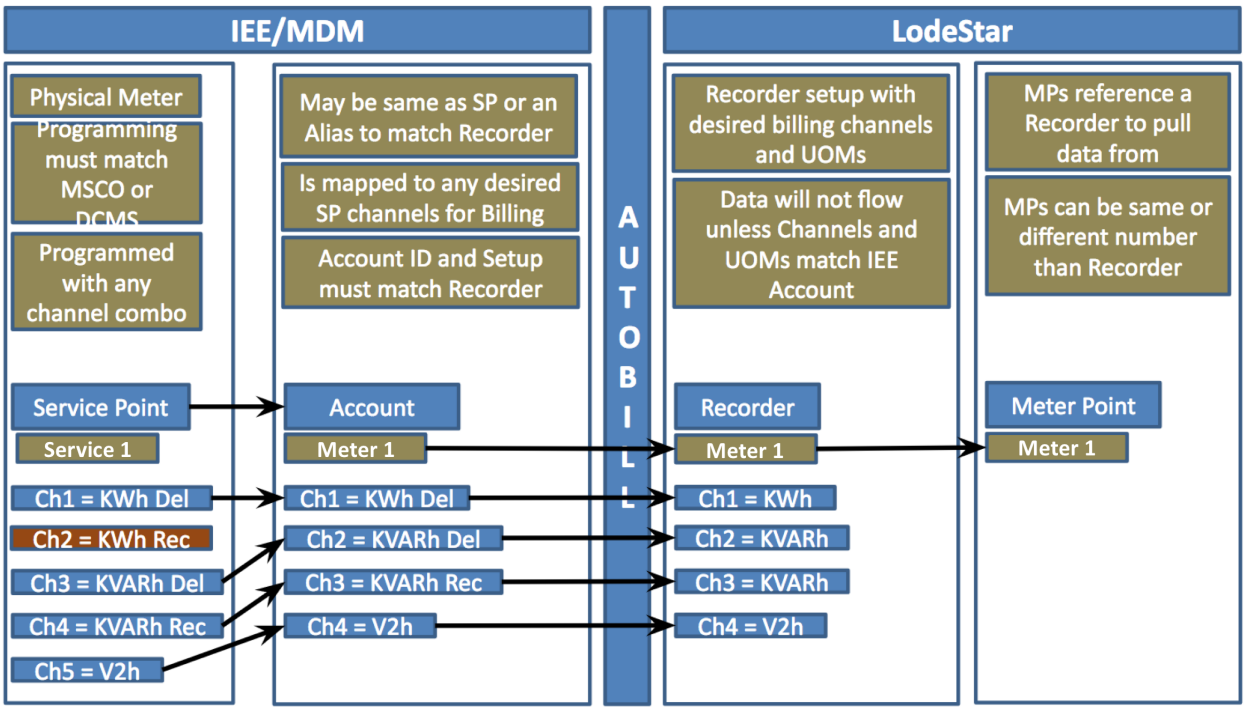


Figure 7.1 Service Point to Meter Point Relationship

7.4.2 Prediction Models

7.4.2.1 Autoregressive Integrated Moving Averages (ARIMA)

The forecasting method we chose to use in initial time series analysis is the ARIMA model. Autoregressive integrated moving average models is one of the most common methods used in time series forecasting, a form of regression analysis with the capability to predict future behavior. The ARIMA model uses previous data to fit a linear equation used in forecasting, in both stationary and non-stationary time series data. When the data is non-stationary, the time series is stationarized through differencing to conform to the requirements of the ARMA model, where the differencing is responsible for the "integrated" aspect of the ARIMA model [130] [131]. Whether the data are seasonal or not determines how many differences are performed on the series to convert it to stationarity. Evaluating autocorrelation is needed to address relative variability in data, which is executed by calculating the differences of one observation to the prior observation, to identify any residual patterns [132]. The "au-

toregressive" (AR) and "moving average" (MA) parts of the acronym stem from amounts of lag, from the differenced series and of the forecast errors, respectively. Components of the ARIMA model include random-walk and trend models, autoregressive models, and exponential smoothing models.

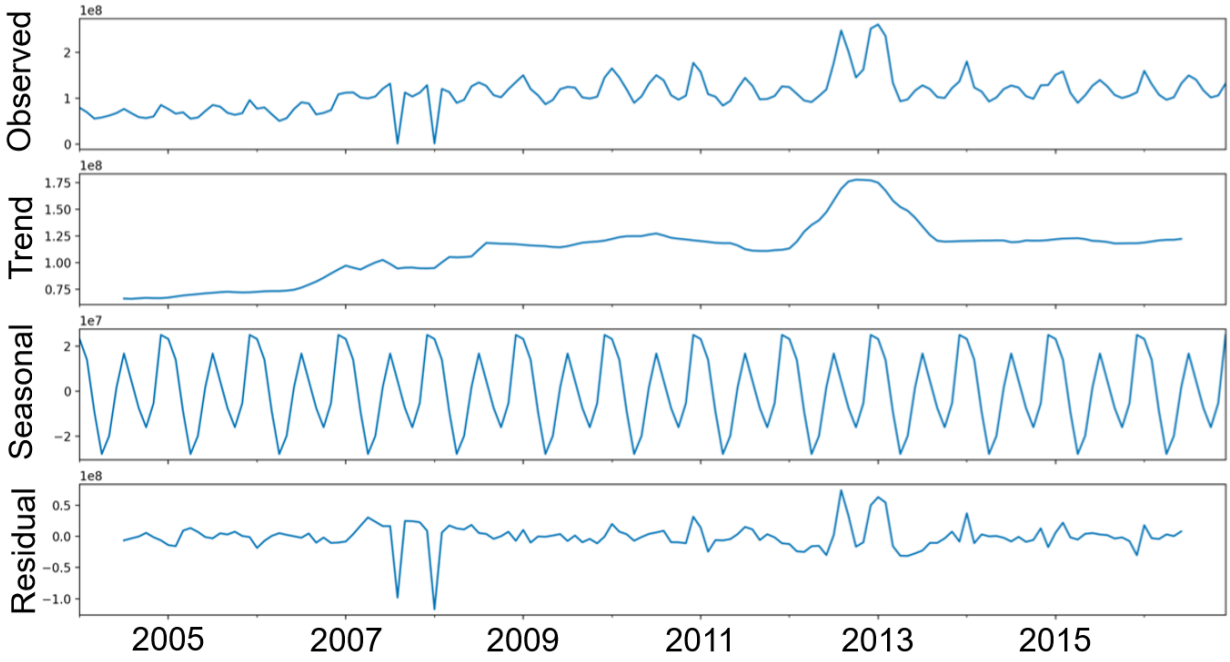


Figure 7.2 ARIMA's breakdown of time series data

Figure 7.2 demonstrates the graphical output after an ARIMA analysis has been performed, showing the observed, trend, seasonal, and residual components of the time series data. Anomaly detection with ARIMA was done with utilization of kWh observed data and residual values.

7.4.2.2 Long Short-Term Memory (LSTM)

Artificial neural networks (ANN) are structured to behave like biological neural networks, analyzing information through networks of computational units called neurons set in layers

[133]. RNNs follow the ANN architecture with analysis of current values and previous values, explaining the term "recurrent" in the name.

Conventional recurrent neural networks rely on either a back-propagation through time [118], or a real-time recurrent learning [134] algorithm, for which we deployed the special RNN architecture LSTM to resolve issues of vanishing/exploding gradient [120]. LSTM is a type of RNN designed to model temporal sequences alongside deep dependencies. It utilizes current input while remembering previous input. LSTM networks use 4 processing units, most commonly composed of a memory cell, an input gate, an output gate, and a forget gate; it is through the different combinations of gates that the neurons decide what values to store and when to allow retrieval, alterations, and removal of information used in training and testing [133].

Table 7.1 Parameters of LSTM Model

Layer	# of Neurons	Dropout	Return Sequence
LSTM 1	60	0.2	TRUE
LSTM 2	30	0.2	TRUE
Output	1	-	-

Table 7.1 shows the parameters of each LSTM layers. The dropout describes the percentage of randomly omitted feature detectors on each training case, which is utilized to reduce overfitting [135]. Mean Squared Error was used as the loss function. Defining the return sequence attribute to TRUE enables LSTM layers to provide an output at each iteration.

7.4.2.3 ARIMA and LSTM Model (Combination Method)

Our proposed method for a new predictive model is a combination of both the ARIMA and LSTM models. To take advantage of ARIMA's ability to address nonlinear components of data as well as LSTM's RNN architecture, we created a model that runs the data through ARIMA to output anomalies that are then run through LSTM to further to minimize false

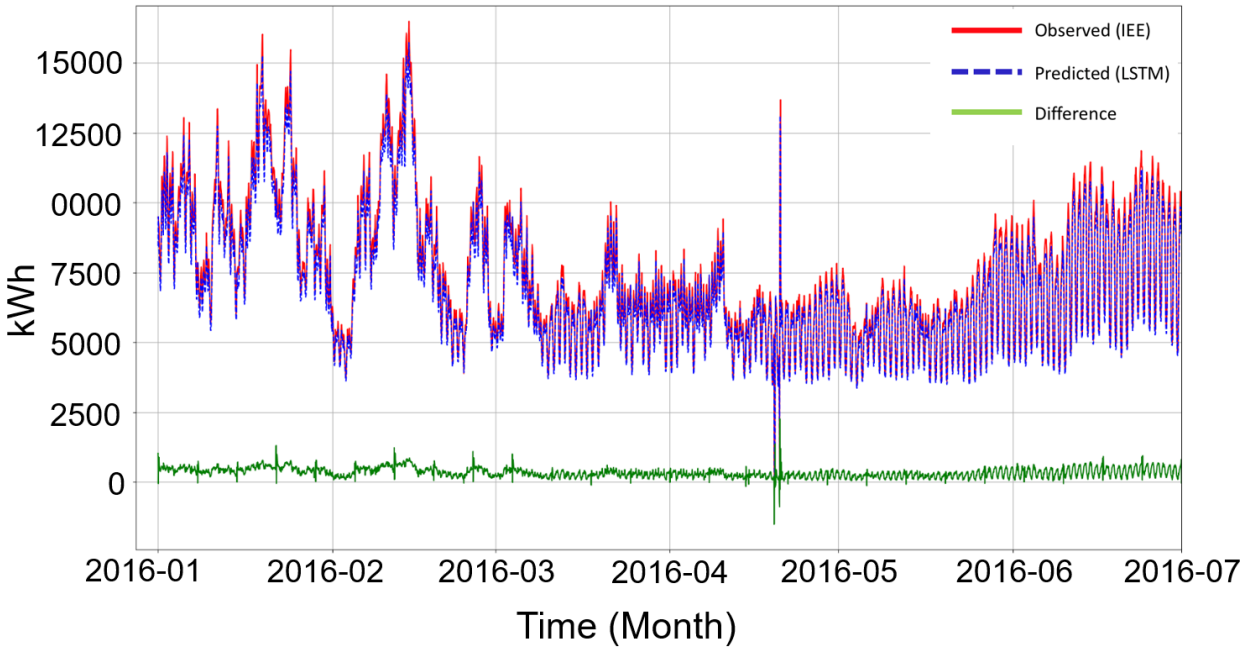


Figure 7.3 Example of LSTM

detections and increase the effectiveness in anomaly detection.

7.5 Results

7.5.1 Model Outputs

Using data from 2014-2016, we performed anomaly detection analysis with ARIMA, LSTM, and the Combination method. Each model analysis was performed using kWh data, with LSTM also including weather data, from 2014-2016. The machine learning models were trained over two years of data and tested for comparison over data from 2016.

ARIMA flagged residual values of the time series data that went above or below the predetermined threshold of 200kWh. An example of the anomaly detection by a single meter with a residue passing this threshold occurred on the date of June 13, 2016 and is shown in Figure 7.4 below.

LSTM found instances where there were significantly large differences between actual and

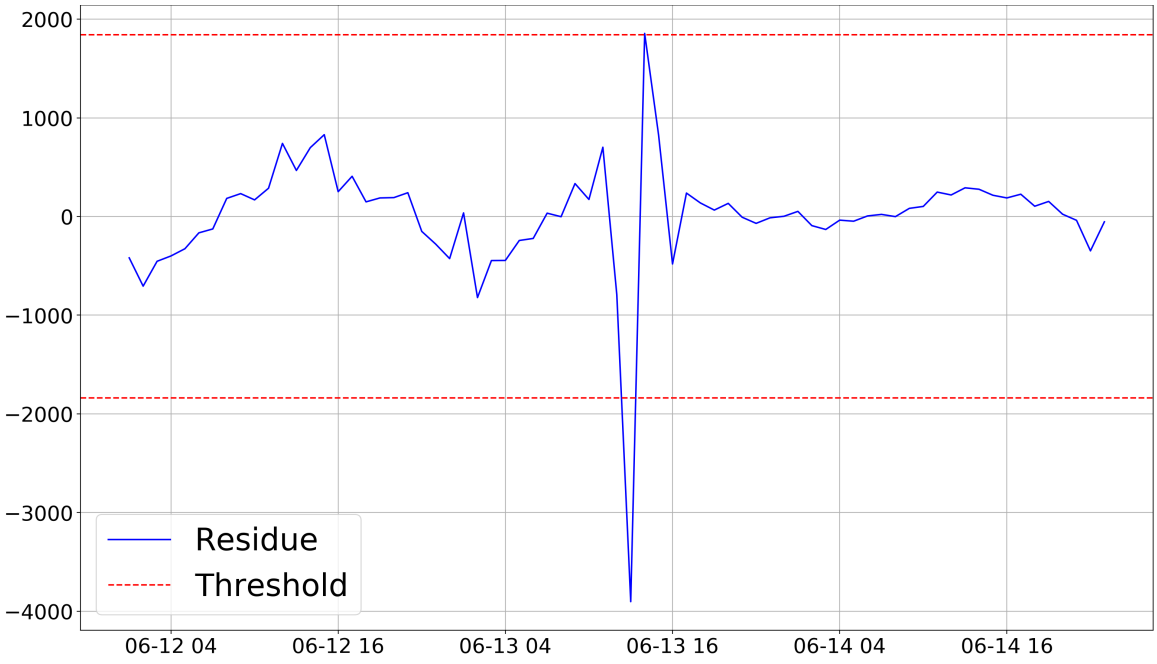


Figure 7.4 Example of ARIMA

predicted values. An example of this anomaly detection by LSTM is shown below in Figure 7.5 and is based on the same meter and time frame that was displayed in Figure 7.4.

The Combination Method predicted potential anomalies by matching the results from ARIMA and LSTM, as true anomalies were detected by the combination of the two methods. Above we showed the results of a single meter on a single day, now we are looking at all meters on a single day to compare the results of the predictive models ability to detect anomalies. The results of the 30 meters for one selected day (June 13, 2016) are shown in Tables 7.2, 7.3, 7.4, where the boldface represents instances in which results match TVA's reporting of load transfers. Tables 7.2 and 7.4 refer to the ARIMA analysis and Combination method, respectively, for the entirety of the selected day, while Table 7.3 refers to the output of LSTM analysis on only a representative portion of the large number of anomalies detected on that day.

According to Table 7.2, the ARIMA model detected 4 anomalies, 3 of which were actually detected by TVA. This represents the ability of ARIMA to detect anomalies and an example

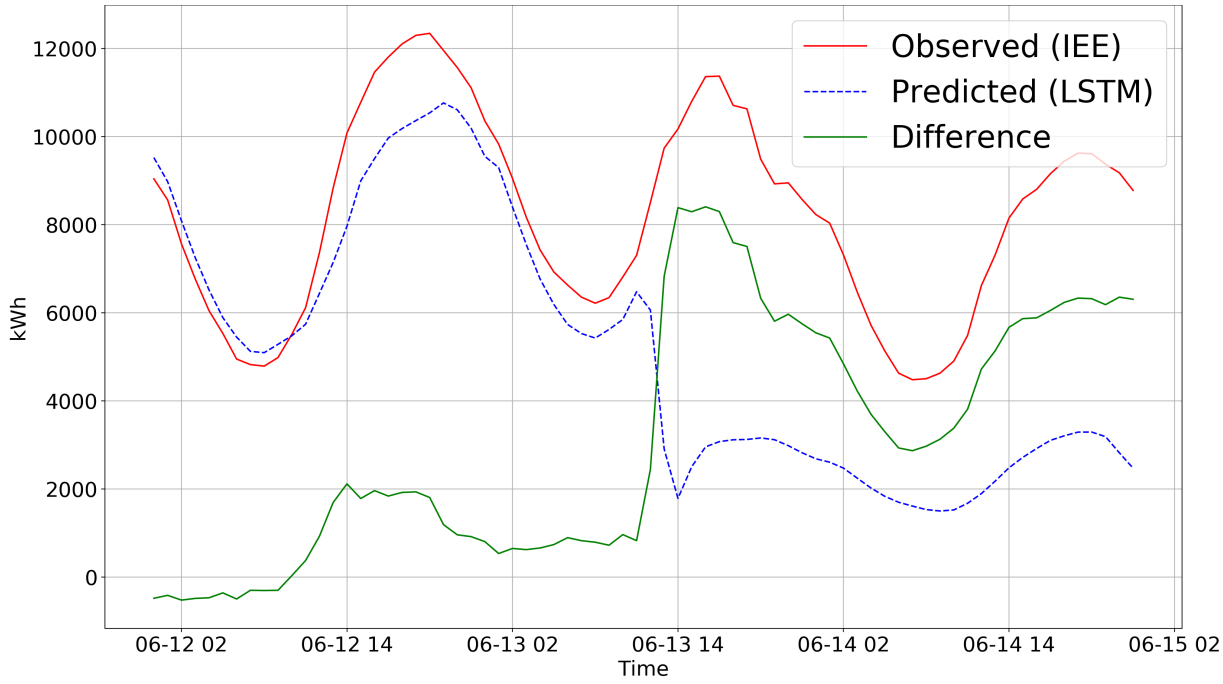


Figure 7.5 Example of LSTM

Table 7.2 Output of ARIMA

Time	Meter	kWh	Residue
06/13/2016 10:00	Meter 4	5566	1063
06/13/2016 12:00	**Meter 1**	**15541**	**1063**
06/13/2016 13:00	**Meter 6**	**1589**	**-3903**
06/13/2016 13:00	**Meter 5**	**12413**	**4648**

where it displays oversensitivity.

Table 7.3 displays the output of LSTM for only one hour of the day, showing 6 anomalies detected, with 3 falsely detected. The oversensitivity and low accuracy of LSTM is indicated by the 3 false detections in only one hour while accumulating a total of 11 falsely detected meters for that day as a whole.

Table 7.4 shows the results of the Combined Method which showed no false detections

Table 7.3 Output of LSTM

Time	Meter	kWh	Predicted kWh	Difference
...	...	...	...	...
06/13/2016 12:00	Meter 2	11561	11292	269
06/13/2016 12:00	Meter 7	4574	4354	220
06/13/2016 12:00	**Meter 1**	**15541**	**15783**	**-242**
06/13/2016 13:00	**Meter 5**	**12413**	**13185**	**-772**
06/13/2016 13:00	Meter 3	8630	8408	222
06/13/2016 13:00	**Meter 6**	**1589**	**2908**	**-1319**
...	...	...	...	...

Table 7.4 Output of Combined Method

Time	Meter	kWh	Predicted kWh	Difference
06/13/2016 12:00	**Meter 1**	**15541**	**15783**	**-242**
06/13/2016 13:00	**Meter 5**	**12413**	**13185**	**-772**
06/13/2016 13:00	**Meter 6**	**1589**	**2908**	**-1319**

and detected all of the true anomalies, indicating the strength of the Combined Method model.

7.5.2 Model Performance

In order to determine the extent of the performance of each model, all the reported dates for 2016 were used as parameters for each model.

Each model was evaluated as a comparison of the meters detected as anomalies by the model to the TVA-provided meter report of anomalies. Table 7.5 below represents a confusion matrix for the results of this comparison. The True Negative (TN) values represent the number of meters that were not detected as anomalies by the model nor indicated as an anomaly by TVA. The False Negative (FN) values represent the number of meters not detected by the model but indicated as an anomaly by TVA. The True Positive (TP) values

represent the number of meters labeled an anomaly by both the model and TVA. Lastly, the False Positive (FP) values represent meters that the model detected as an anomaly but were not identified by TVA.

Table 7.5 Confusion Matrix for Each Predictive Model

Anomaly	ARIMA	LSTM	Combination Method
TN	856	555	956
FN	17	24	29
TP	85	81	75
FP	162	460	60

The following equations were used to calculate the accuracy (7.1), True Positive Rate (7.2), False Positive Rate (7.3), and specificity (7.4) from variables of the Confusion Matrix (Table 7.5) for the purpose of evaluating the performance of each prediction model. :

$$Accuracy = \frac{TP + TN}{total} \tag{7.1}$$

$$TruePositiveRate = \frac{TP}{FN + TP} \tag{7.2}$$

$$FalsePositiveRate = \frac{FP}{TN + FP} \tag{7.3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7.4}$$

Table 7.6 Performance Comparison of Predictive Models

Anomaly	ARIMA	LSTM	Combination Method
Accuracy	0.840	0.568	0.921
True Positive Rate	0.833	0.771	0.690
False Positive Rate	0.159	0.453	0.061
Specificity	0.841	0.547	0.939

According to Table 7.6, the Combination method had the highest accuracy and specificity with the lowest False Positive Rate out of the three methods. The only category where the Combination Method did not outperform LSTM and ARIMA was in the True Positive Rate category, where a higher True Positive Rate is desired. As suggested by the output of the LSTM meters, shown in Table 7.3, LSTM had the lowest accuracy rate, highest False Positive Rate, and lowest specificity of the three models.

7.6 Conclusion

With the use of Python and its aptitude in time series analysis and machine learning, we show an application in the detection of anomalies with a focus on meter load transfers. The machine learning algorithm created with both ARIMA and LSTM models proves to identify anomalies in the energy delivery system with the added capability to provide the time of the incident. The proposed predictive model of combining the ARIMA and LSTM methods resulted in the highest accuracy out of all the model performances. Additionally, by providing the minimal amount of false anomalies detected, there are fewer meters that need to be reviewed manually. This method of meter review would lessen the both the financial cost and time spent on tedious data analysis and recovery in the power utilities industry, enhancing the quality of meter data by use of an automated process. Without the implementation of our machine learning model, over 1440 data points have to be reviewed with reliance on manual detection – using our model, only select meters would need to be reviewed.

CHAPTER 8
CONCLUSION

Two important smart city initiatives: Smart Health and Smart Energy were investigated with the goal of "advance the quality of life through technology and data science" in mind. Various types of statistical analysis and machine learning approaches were considered to accommodate different datasets studied in this dissertation.

Based on the hospital discharge data system provided by the Tennessee Department of Health, two studies were investigated: statistical analysis based smart health and machine learning based smart health. As one of the major research topic in computer science, these two studies can be broken further down: a white-box approach vs. a black-box approach in health settings. White-box models such as logistic regression allows for clear interpretation on how they behave and it's easy to explore how different variables react and influence each other. On the other hand, black-box models rely on complex mathematical computations to analyze data and produce output in high performance while not knowing what is actually happening within the model. The choice between white-box models or black-box models becomes complicated especially in the health settings. Thus, exploring both options gives a general idea of the strengths and weaknesses of each proposed models, and physicians or clinicians could take this information to further assist their patients.

Based on the data from the Centers for Medicare and Medicaid Services, hospital readmission rate was explored to further validate the clinical practicality of the proposed smart health solutions. A threshold value was introduced to classify and separate the value between the probability of facility discharge versus home discharge. However, this threshold value was

solely based on the probability plot without the clinical input from physicians. In the future, this system can be adjusted with the help from physicians to implement a dynamic threshold selection process that could suggest results that are more clinically significant. Alongside of medical claims data, a text-based data also provides clinically significant insights about the patients and their outcomes. While analyzing medical claims data are crucial in completing achieving the goal of smart health initiatives, machine learning models combined with natural language processing (NLP) techniques are becoming a trend among modern smart city applications. In this dissertation, a patient-specific clinical trial matching system was developed using the clinical trial data provided by the ClinicalTrials.gov. Based on the proposed solution, clusters containing patient-specific keywords are returned to the user to quickly return information associated with the patient as well as the recommended trials for the patient. With this system, clinicians can save tremendous time by not having to look through every available trials that can be matched to the patient. However, this system could improve further by adding a dashboard or a type of visualization for the clinicians to quickly view the clusters. Also, comparing the result with handpicked trials selected by clinicians will also help in refinement and validation of the proposed solution.

In addition to the smart health initiatives, two smart energy projects were investigated to further enhance the energy efficiency in both residential and industrial sectors within the scope of the smart city. The first project used the residential smart meter data provided Pecan Street which includes smart meter records for total energy usage and air conditioning energy usage. Two popular models for time series data were considered: ARIMA and LSTM. Both models demonstrated high accuracy with minimal errors. While ARIMA achieved high accuracy in short-term forecasting, LSTM showed exceptional performance in long-term forecasting with the use of large dataset.

For industrial sector, anomaly detection algorithm was develop to assist TVA in finding anomalies within their meter data. Three different approaches were considered to assess the strength of each proposed models. Stand-alone ARIMA model yielded high accuracy results

(accuracy of 0.840) compared to stand-alone LSTM model (accuracy of 0.568). However, false positive rate was still high despite the high accuracy. Therefore, a combination model was proposed to address high false positive rate among stand-alone models while maintaining high prediction accuracy. Out of the three models, the combination model had the highest accuracy (accuracy of 0.921). The combination model maximized the strengths from both models in finding the difference between actual value and the predicted value. First, ARIMA is computationally less expensive than LSTM therefore, it quickly yields prediction results. However, some prediction errors can be reduced further by comparing the results with the output from the LSTM model. Considering the size of the provided dataset, LSTM could be more helpful in the future when the dataset size increases (historical & current data) since LSTM is more suitable for large datasets.

Smart city applications are critical in city planning to alter the modern city into more efficient and sustainable smart city. Proposed smart city initiatives can be further improved with the help of domain experts in validating the effectiveness and the adaptability.

REFERENCES

[1] T. Bakıcı, E. Almirall, and J. Wareham, "A smart city initiative: the case of barcelona," *Journal of the knowledge economy*, vol. 4, no. 2, pp. 135–148, 2013.

[2] A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. A. Pérez-Martínez, R. Di Pietro, D. N. Perrea *et al.*, "Smart health: A context-aware health paradigm within smart cities," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 74–81, 2014.

[3] T. Nam and T. A. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," in *Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times*, 2011, pp. 282–291.

[4] I. Dincer and C. Acar, "Smart energy systems for a sustainable future," *Applied Energy*, vol. 194, pp. 225–235, 2017.

[5] H. Lund, P. A. Østergaard, D. Connolly, and B. V. Mathiesen, "Smart energy and smart energy systems," *Energy*, vol. 137, pp. 556–565, 2017.

[6] M. L. Clark and T. Gropen, "Advances in the stroke system of care," *Current treatment options in cardiovascular medicine*, vol. 17, no. 1, pp. 1–13, 2015.

[7] "Stroke facts," http://www.cdc.gov/stroke/facts.htm, March 2015.

[8] "Prevalence of stroke," https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6120a5.htm, May 2012.

[9] "Tennessee stroke registry report," 2014.

[10] L. Chan, M. E. Sandel, A. M. Jette, J. Appelman, D. E. Brandt, P. Cheng, M. TeSelle, R. Delmonico, J. F. Terdiman, and E. K. Rasch, "Does postacute care site matter? a longitudinal study assessing functional recovery after a stroke," *Archives of physical medicine and rehabilitation*, vol. 94, no. 4, pp. 622–629, 2013.

[11] J. A. Luker, J. Bernhardt, K. A. Grimmer, and I. Edwards, "A qualitative exploration of discharge destination as an outcome or a driver of acute stroke care," *BMC health services research*, vol. 14, no. 1, p. 1, 2014.

[12] T.-A. Nguyen, A. Page, A. Aggarwal, and P. Henke, "Social determinants of discharge destination for patients after stroke with low admission fim instrument scores," *Archives of physical medicine and rehabilitation*, vol. 88, no. 6, pp. 740–744, 2007.

[13] A. J. Kind, M. A. Smith, J.-I. Liou, N. Pandhi, J. R. Frytak, and M. D. Finch, "Discharge destination's effect on bounce-back risk in black, white, and hispanic acute ischemic stroke patients," *Archives of physical medicine and rehabilitation*, vol. 91, no. 2, pp. 189–195, 2010.

[14] J. K. Freburger, G. M. Holmes, L.-J. E. Ku, M. P. Cutchin, K. Heatwole-Shank, and L. J. Edwards, "Disparities in postacute rehabilitation care for stroke: an analysis of the state inpatient databases," *Archives of physical medicine and rehabilitation*, vol. 92, no. 8, pp. 1220–1229, 2011.

[15] Y. Béjot, O. Troisgros, V. Gremeaux, B. Lucas, A. Jacquin, C. Khoumri, C. Aboa-Eboulé, C. Benaïm, J.-M. Casillas, and M. Giroud, "Poststroke disposition and associated factors in a population-based study the dijon stroke registry," *Stroke*, vol. 43, no. 8, pp. 2071–2077, 2012.

[16] B. N. Jaja, G. Saposnik, R. Nisenbaum, B. W. Lo, T. A. Schweizer, K. E. Thorpe, and R. L. Macdonald, "Racial/ethnic differences in inpatient mortality and use of institutional postacute care following subarachnoid hemorrhage: clinical article," *Journal of neurosurgery*, vol. 119, no. 6, pp. 1627–1632, 2013.

[17] Y. Xian, R. G. Holloway, E. E. Smith, L. H. Schwamm, M. J. Reeves, D. L. Bhatt, P. J. Schulte, M. Cox, D. M. Olson, A. F. Hernandez *et al.*, "Racial/ethnic differences in process of care and outcomes among patients hospitalized with intracerebral hemorrhage," *Stroke*, vol. 45, no. 11, pp. 3243–3250, 2014.

[18] K. Van der Cruyssen, L. Vereeck, W. Saeys, and R. Remmen, "Prognostic factors for discharge destination after acute stroke: a comprehensive literature review," *Disability and rehabilitation*, vol. 37, no. 14, pp. 1214–1227, 2015.

[19] H.-P. Tseng, F.-J. Lin, P.-T. Chen, C.-H. Mou, S.-P. Lee, C.-Y. Chang, A.-C. Chen, C.-H. Liu, C.-H. Yeh, S.-Y. Tsai *et al.*, "Derivation and validation of a discharge disposition predicting model after acute stroke," *Journal of Stroke and Cerebrovascular Diseases*, vol. 24, no. 6, pp. 1179–1186, 2015.

[20] V. Q. Nguyen, J. PrvuBettger, T. Guerrier, M. A. Hirsch, J. G. Thomas, T. M. Pugh, and C. F. Rhoads, "Factors associated with discharge to home versus discharge to institutional care after inpatient stroke rehabilitation," *Archives of physical medicine and rehabilitation*, vol. 96, no. 7, pp. 1297–1303, 2015.

[21] C. L. Bell, A. Z. LaCroix, M. Desai, H. Hedlin, S. R. Rapp, C. Cene, J. Savla, T. Shippee, S. Wassertheil-Smoller, M. L. Stefanick *et al.*, "Factors associated with nursing home admission after stroke in older women," *Journal of Stroke and Cerebrovascular Diseases*, vol. 24, no. 10, pp. 2329–2337, 2015.

[22] R. D. Dutrieux, M. van Eijk, M. L. van Mierlo, C. M. van Heugten, J. Visser-Meily, and W. P. Achterberg, "Discharge home after acute stroke: Differences between older and younger patients," *Journal of rehabilitation medicine*, vol. 48, no. 1, pp. 14–18, 2016.

[23] M. Mees, J. Klein, L. Yperzeele, P. Vanacker, and P. Cras, "Predicting discharge destination after stroke: A systematic review," *Clinical neurology and neurosurgery*, vol. 142, pp. 15–21, 2016.

[24] J. Stein, J. P. Bettger, A. Sicklick, R. Hedeman, Z. Magdon-Ismail, and L. H. Schwamm, "Use of a standardized assessment to predict rehabilitation care after acute stroke," *Archives of physical medicine and rehabilitation*, vol. 96, no. 2, pp. 210–217, 2015.

[25] *Hospital Discharge Data System User Manual*, Division of Health Statistics, Tennessee Department of Health, 2011.

[26] L. M. Sullivan, J. M. Massaro, and R. B. D'Agostino, "Presentation of multivariate data for clinical use: The framingham study risk score functions," *Statistics in medicine*, vol. 23, no. 10, pp. 1631–1660, 2004.

[27] C. Wu, E. L. Hannan, G. Walford, J. A. Ambrose, D. R. Holmes, S. B. King, L. T. Clark, S. Katz, S. Sharma, and R. H. Jones, "A risk score to predict in-hospital mortality for percutaneous coronary interventions," *Journal of the American College of Cardiology*, vol. 47, no. 3, pp. 654–660, 2006.

[28] M. Kelly-Hayes, P. A. Wolf, W. B. Kannel, P. Sytkowski, R. B. D'Agostino, and G. E. Gresham, "Factors influencing survival and need for institutionalization following stroke: the framingham study." *Archives of physical medicine and rehabilitation*, vol. 69, no. 6, pp. 415–418, 1988.

[29] P. d. Pablo, E. Losina, C. B. Phillips, A. H. Fossel, N. Mahomed, E. A. Lingard, and J. N Katz, "Determinants of discharge destination following elective total hip replacement," *Arthritis Care & Research*, vol. 51, no. 6, pp. 1009–1017, 2004.

[30] J. Osborne, C. D. Langefeld, C. J. Moomaw, K. N. Sheth, D. Y. Hwang, M. L. Flaherty, A. Vashkevich, J. Gohs, and D. Woo, "Abstract ns20: Discharge disposition after intracerebral hemorrhage," *Stroke*, vol. 46, no. Suppl 1, pp. ANS20–ANS20, 2015.

[31] S.-Y. Wang, Y. Zhao, and X.-Y. Zang, "Continuing care for older patients during the transitional period," *Chinese Nursing Research*, vol. 1, pp. 5–13, 2014.

[32] K. K. Andersen, T. S. Olsen, C. Dehlendorff, and L. P. Kammersgaard, "Hemorrhagic and ischemic strokes compared," *Stroke*, vol. 40, no. 6, pp. 2068–2072, 2009.

[33] J. P. Bettger, X. Zhao, C. Bushnell, L. Zimmer, W. Pan, L. S. Williams, and E. D. Peterson, "The association between socioeconomic status and disability after stroke: findings from the adherence evaluation after ischemic stroke longitudinal (avail) registry," *BMC public health*, vol. 14, no. 1, p. 281, 2014.

[34] M. G. Stineman, P. L. Kwong, B. E. Bates, J. E. Kurichi, D. C. Ripley, and D. Xie, "Development and validation of a discharge planning index for achieving home discharge after hospitalization for acute stroke among those who received rehabilitation services," *American Journal of Physical Medicine & Rehabilitation*, vol. 93, no. 3, pp. 217–230, 2014.

[35] N. El Husseini, G. C. Fonarow, E. E. Smith, C. Ju, L. H. Schwamm, A. F. Hernandez, P. J. Schulte, Y. Xian, and L. B. Goldstein, "Renal dysfunction is associated with poststroke discharge disposition and in-hospital mortality," *Stroke*, pp. STROKEAHA–116, 2016.

[36] "Stroke facts," http://www.cdc.gov/stroke/facts.htm.

[37] "Hypertension among adults in the united states: National health and nutrition examination survey, 2011-2012," https://www.cdc.gov/nchs/products/databriefs/db133.htm, October 2013.

[38] "Vital signs: avoidable deaths from heart disease, stroke, and hypertensive disease-united states, 2001-2010." pp. 721–727, 2013.

[39] D. Ouellette, C. Timple, S. Kaplan, S. Rosenberg, and E. Rosario, "Predicting discharge destination with admission outcome scores in stroke patients," *NeuroRehabilitation*, vol. 37, no. 2, pp. 173–179, 2015.

[40] M. T. Fox, M. Persaud, I. Maimets, D. Brooks, K. O'Brien, and D. Tregunno, "Effectiveness of early discharge planning in acutely ill or injured hospitalized older adults: a systematic review and meta-analysis," *BMC geriatrics*, vol. 13, no. 1, p. 70, 2013.

[41] R. Beech, A. G. Rudd, K. Tilling, and C. D. Wolfe, "Economic consequences of early inpatient discharge to community-based rehabilitation for stroke in an inner-london teaching hospital," *Stroke*, vol. 30, no. 4, pp. 729–735, 1999.

[42] M. À. Mas and M. Inzitari, "A critical review of early supported discharge for stroke patients: from evidence to implementation into practice," *International Journal of Stroke*, vol. 10, no. 1, pp. 7–12, 2015.

[43] M. J. Meyer, S. Pereira, A. McClure, R. Teasell, A. Thind, J. Koval, M. Richardson, and M. Speechley, "A systematic review of studies reporting multivariable models to predict functional outcomes after post-stroke inpatient rehabilitation," *Disability and rehabilitation*, vol. 37, no. 15, pp. 1316–1323, 2015.

[44] J. P. Bettger, L. Thomas, L. Liang, Y. Xian, C. D. Bushnell, J. L. Saver, G. C. Fonarow, and E. D. Peterson, "Hospital variation in functional recovery after stroke," *Circulation: Cardiovascular Quality and Outcomes*, vol. 10, no. 1, p. e002391, 2017.

[45] M. Alcusky, C. M. Ulbricht, and K. L. Lapane, "Postacute care setting, facility characteristics, and poststroke outcomes: a systematic review," *Archives of physical medicine and rehabilitation*, vol. 99, no. 6, pp. 1124–1140, 2018.

[46] J. S. Cho, Z. Hu, N. Fell, G. W. Heath, R. Qayyum, and M. Sartipi, "Hospital discharge disposition of stroke patients in tennessee." *Southern medical journal*, vol. 110, no. 9, pp. 594–600, 2017.

[47] M. Bailey, A. Weiss, M. Barrett, and H. Jiang, "Characteristics of 30-day all-cause hospital readmissions, 2010-2016: Statistical brief #248," 2019.

[48] S. F. Jencks, M. V. Williams, and E. A. Coleman, "Rehospitalizations among patients in the medicare fee-for-service program," *New England Journal of Medicine*, vol. 360, no. 14, pp. 1418–1428, 2009.

[49] CMS, "Hospital readmissions reduction program (hrrp)," 2012. [Online]. Available: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html

[50] C. K. McIlvennan, Z. J. Eapen, and L. A. Allen, "Hospital readmissions reduction program," *Circulation*, vol. 131, no. 20, pp. 1796–1803, 2015.

[51] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, "Risk prediction models for hospital readmission: a systematic review," *Jama*, vol. 306, no. 15, pp. 1688–1698, 2011.

[52] I. Shams, S. Ajorlou, and K. Yang, "A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or copd," *Health care management science*, vol. 18, no. 1, pp. 19–34, 2015.

[53] J. Collins, I. M. Abbass, R. Harvey, B. Suehs, C. Uribe, J. Bouchard, T. Prewitt, T. DeLuzio, and E. Allen, "Predictors of all-cause 30 day readmission among medicare patients with type 2 diabetes," *Current medical research and opinion*, vol. 33, no. 8, pp. 1517–1523, 2017.

[54] E. Benjamin, P. Muntner, A. Alonso, M. Bittencourt, C. Callaway, A. Carson, A. Chamberlain, A. Chang, S. Cheng, S. Das *et al.*, "Heart disease and stroke statistics-2019 update," *Circulation*, vol. 139, no. 10, 2019.

[55] A. M. Nouh, L. McCormick, J. Modak, G. Fortunato, and I. Staff, "High mortality among 30-day readmission after stroke: Predictors and etiologies of readmission," *Frontiers in neurology*, vol. 8, p. 632, 2017.

[56] J. H. Lichtman, E. C. Leifheit-Limson, S. B. Jones, Y. Wang, and L. B. Goldstein, "Preventable readmissions within 30 days of ischemic stroke among medicare beneficiaries," *Stroke*, vol. 44, no. 12, pp. 3429–3435, 2013.

[57] W. Zhong, N. Geng, P. Wang, Z. Li, and L. Cao, "Prevalence, causes and risk factors of hospital readmissions after acute stroke and transient ischemic attack: a systematic review and meta-analysis," *Neurological Sciences*, vol. 37, no. 8, pp. 1195–1202, 2016.

[58] M. J. Hall, S. Levant, and C. J. DeFrances, "Hospitalization for stroke in us hospitals, 1989–2009," *Diabetes*, vol. 18, no. 23, p. 23, 2012.

[59] A. K. Boehme, C. Esenwa, and M. S. Elkind, "Stroke risk factors, genetics, and prevention," *Circulation research*, vol. 120, no. 3, pp. 472–495, 2017.

[60] A. Rao, E. Barrow, S. Vuik, A. Darzi, and P. Aylin, "Systematic review of hospital readmissions in stroke patients," *Stroke research and treatment*, vol. 2016, 2016.

[61] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[62] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[63] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics*, vol. 35, no. 5-6, pp. 352–359, 2002.

[64] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[65] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[66] J. Cho, Z. Hu, and M. Sartipi, "Post-stroke discharge disposition prediction using deep learning," in *SoutheastCon 2017*. IEEE, 2017, pp. 1–2.

[67] O. Blatchford, W. R. Murray, and M. Blatchford, "A risk score to predict need for treatment for uppergastrointestinal haemorrhage," *The Lancet*, vol. 356, no. 9238, pp. 1318–1321, 2000.

[68] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[69] Singh, Chandan, Murdoch, W. James, Yu, and Bin, "Hierarchical interpretations for neural network predictions," Jan 2019. [Online]. Available: https://arxiv.org/abs/1806.05337

[70] R. E. Wright, "Logistic regression." *Reading and understanding multivariate statistics*, pp. 217–244, 1995.

[71] I. Barandiaran, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 20, no. 8, pp. 1–22, 1998.

[72] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[73] C. for Disease Control and Prevention, "U.s. cancer statistics data visualization tool," n.d., [Online; accessed 20-Feb-2021].

[74] H. K. Weir, T. D. Thompson, A. Soman, B. Møller, and S. Leadbetter, "The past, present, and future of cancer incidence in the united states: 1975 through 2020," *Cancer*, vol. 121, no. 11, pp. 1827–1837, 2015.

[75] J. M. Unger, E. Cook, E. Tai, and A. Bleyer, "The role of clinical trial participation in cancer research: barriers, evidence, and strategies," *American Society of Clinical Oncology Educational Book*, vol. 36, pp. 185–198, 2016.

[76] T. Hao, A. Rusanov, M. R. Boland, and C. Weng, "Clustering clinical trials with similar eligibility criteria features," *Journal of biomedical informatics*, vol. 52, pp. 112–120, 2014.

[77] M. R. Boland, R. Miotto, J. Gao, and C. Weng, "Feasibility of feature-based indexing, clustering, and search of clinical trials: A case study of breast cancer trials from clinicaltrials. gov," *Methods of information in medicine*, vol. 52, no. 5, p. 382, 2013.

[78] Y. Ni, J. Wright, J. Perentesis, T. Lingren, L. Deleger, M. Kaiser, I. Kohane, and I. Solti, "Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients," *BMC medical informatics and decision making*, vol. 15, no. 1, pp. 1–10, 2015.

[79] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, no. 6, 2013, pp. 380–384.

[80] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[81] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 407–416.

[82] U. D. of Energy, "Air conditioning," n.d., [Online; accessed 10-June-2019].

[83] EIA, "Household energy use in texas: A closer look at residential energy consumption," Energy Information Administration, Tech. Rep., 2009.

[84] L. W. Davis and P. J. Gertler, "Contribution of air conditioning adoption to future energy use under global warming," *Proceedings of the National Academy of Sciences*, vol. 112, no. 19, pp. 5962–5967, 2015.

[85] K. Perez, W. Cole, M. Baldea, and T. Edgar, "Meters to models: Using smart meter data to predict and control home energy use," in *ACEEE 2014 Summer Study on Energy Efficiency in Buildings*. ACEEE, 2014.

[86] J. Zuo, S. Pullen, J. Palmer, H. Bennetts, N. Chileshe, and T. Ma, "Impacts of heat waves and corresponding measures: a review," *Journal of Cleaner Production*, vol. 92, pp. 1–12, 2015.

[87] A. Sakka, M. Santamouris, I. Livada, F. Nicol, and M. Wilson, "On the thermal performance of low income housing during heat waves," *Energy and Buildings*, vol. 49, pp. 69–77, 2012.

[88] W. Zhang, J. Lian, C.-Y. Chang, and K. Kalsi, "Aggregated modeling and control of air conditioning loads for demand response," *IEEE transactions on power systems*, vol. 28, no. 4, pp. 4655–4664, 2013.

[89] P. Siano, "Demand response and smart grids-a survey," *Renewable and sustainable energy reviews*, vol. 30, pp. 461–478, 2014.

[90] D. Egarter, V. P. Bhuvana, and W. Elmenreich, "Paldi: Online load disaggregation via particle filtering," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 2, pp. 467–477, 2014.

[91] Y. Li, Z. Peng, J. Huang, Z. Zhang, and J. H. Son, "Energy disaggregation via hierarchical factorial hmm," in *Proceedings of the 2nd International Workshop on Non-Intrusive Load Monitoring*, 2014.

[92] N. F. Esa, M. P. Abdullah, and M. Y. Hassan, "A review disaggregation method in non-intrusive appliance load monitoring," *Renewable and Sustainable Energy Reviews*, vol. 66, pp. 163–173, 2016.

[93] G. Gross and F. D. Galiana, "Short-term load forecasting," *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1558–1573, 1987.

[94] A. Meyler, G. Kenny, and T. Quinn, "Forecasting irish inflation using arima models," *Munich Personal RePEc Archive*, vol. 1998, no. 3, pp. 1–48, 1998.

[95] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[96] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, 2015.

[97] J. M. Jiménez, L. Stokes, C. Moss, Q. Yang, and V. N. Livina, "Modelling energy demand response using long short-term memory neural networks," *Energy Efficiency*, vol. 13, no. 6, pp. 1263–1280, 2020.

[98] H. Son and C. Kim, "A deep learning approach to forecasting monthly demand for residential–sector electricity," *Sustainability*, vol. 12, no. 8, p. 3103, 2020.

[99] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using cnn-lstm neural networks," *Energy*, vol. 182, pp. 72–81, 2019.

[100] S. H. Pramono, M. Rohmatillah, E. Maulana, R. N. Hasanah, and F. Hario, "Deep learning-based short-term load forecasting for supporting demand response program in hybrid energy system," *Energies*, vol. 12, no. 17, p. 3359, 2019.

[101] Z. A. Khan, T. Hussain, A. Ullah, S. Rho, M. Lee, and S. W. Baik, "Towards efficient electricity forecasting in residential and commercial buildings: A novel hybrid cnn with a lstm-ae based framework," *Sensors*, vol. 20, no. 5, p. 1399, 2020.

[102] M. Waseem, Z. Lin, and L. Yang, "Data-driven load forecasting of air conditioners for demand response using levenberg–marquardt algorithm-based ann," *Big Data and Cognitive Computing*, vol. 3, no. 3, p. 36, 2019.

[103] S. Su, Y. Yan, H. Lu, L. Kangping, S. Yujing, W. Fei, L. Liming, and R. Hui, "Non-intrusive load monitoring of air conditioning using low-resolution smart meter data," in *2016 IEEE International Conference on Power System Technology (POWERCON)*. IEEE, 2016, pp. 1–5.

[104] J. Cho, Z. Hu, and M. Sartipi, "Non-intrusive a/c load disaggregation using deep learning," in *2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*. IEEE, 2018, pp. 1–5.

[105] C. Christensen, R. Anderson, S. Horowitz, A. Courtney, and J. Spencer, "Beopt (tm) software for building energy optimization: Features and capabilities," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2006.

[106] L. Setter, E. Smoorenburg, S. Wijesuriya, and P. C. Tabares-Velasco, "Energy and hygrothermal performance of cross laminated timber single-family homes subjected to constant and variable electric rates," *Journal of Building Engineering*, 2019.

[107] J. D. Rhodes, W. H. Gorman, C. R. Upshaw, and M. E. Webber, "Using beopt (energyplus) with energy audits and surveys to predict actual residential energy usage," *Energy and Buildings*, vol. 86, pp. 808–816, 2015.

[108] W. J. Cole, J. D. Rhodes, W. Gorman, K. X. Perez, M. E. Webber, and T. F. Edgar, "Community-scale residential air conditioning control for effective grid management," *Applied Energy*, vol. 130, pp. 428–436, 2014.

[109] K. S. Cetin, M. H. Fathollahzadeh, N. Kunwar, H. Do, and P. C. Tabares-Velasco, "Development and validation of an hvac on/off controller in energyplus for energy simulation of residential and small commercial buildings," *Energy and Buildings*, vol. 183, pp. 467–483, 2019.

[110] A. ASHRAE, "Ashrae guideline 14: Measurement of energy and demand savings," *American Society of Heating, Refrigerating and Air-Conditioning Engineers*, vol. 35, pp. 41–63, 2002.

[111] IECC, "Iecc compliance guide for homes in texas: Code: 2009 international energy conservation code," IECC, Tech. Rep., 2009.

[112] S. Abe, *Support vector machines for pattern classification*. Springer, 2005, vol. 2.

[113] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, p. 1565, 2006.

[114] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data mining techniques for the life sciences*. Springer, 2010, pp. 223–239.

[115] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[116] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[117] R. Nau, "Introduction to arima: nonseasonal models," n.d., [Online; accessed 7-June-2019]. [Online]. Available: https://people.duke.edu/~rnau/411arim.htm

[118] P. J. Werbos *et al.*, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[119] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[120] D. L. Marino, K. Amarasinghe, and M. Manic, "Building energy load forecasting using deep neural networks," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2016, pp. 7046–7051.

[121] T. Zebin, M. Sperrin, N. Peek, and A. J. Casson, "Human activity recognition from inertial sensor time-series using batch normalized deep lstm recurrent networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1–4.

[122] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[123] F. Chollet *et al.*, "Keras," "https://keras.io", 2015.

[124] E. Almeshaiei and H. Soltan, "A methodology for electric power load forecasting," *Alexandria Engineering Journal*, vol. 50, no. 2, pp. 137–144, 2011.

[125] V. Jadhav and V. Ligay, "Forecasting Energy Consumption using Machine Learning." *ResearchGate*, 2016.

[126] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, 2017.

[127] Q. Zhu, Y. Guo, and G. Feng, "Household energy consumption in China: Forecasting with BVAR model up to 2015," in *Computational Sciences and Optimization (CSO), 2012 Fifth International Joint Conference on*. IEEE, 2012, pp. 654–659.

[128] D. Anderson, J.-R. Vlimant, and M. Spiropulu, "An MPI-Based Python Framework for Distributed Training with Keras," *arXiv preprint arXiv:1712.05878*, 2017.

[129] J. B. Diederik P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[130] V. Panasa, R. V. Kumari, G. Ramakrishna, and S. Kaviraju, "Maize Price Forecasting Using Auto Regressive Integrated Moving Average (ARIMA) Model," *Int. J. Curr. Microbiol. App. Sci*, vol. 6, no. 8, pp. 2887–2895, 2017.

[131] R. K. Sharma, "Forecasting Gold price with Box Jenkins Autoregressive Integrated Moving Average Method," *Journal of International Economics*, vol. 7, no. 1, p. 32, 2016.

[132] R. C. Sato, "Disease management with ARIMA model in time series," *Einstein (Sao Paulo)*, vol. 11, no. 1, pp. 128–131, 2013.

[133] D. Janardhanan and E. Barrett, "CPU Workload forecasting of Machines in Data Centers using LSTM Recurrent Neural Networks and ARIMA Models," *ResearchGate*, 2018.

[134] R. J. Williams and D. Zipser, "Gradient-based learning algorithms for recurrent networks and their computational complexity," *Backpropagation: Theory, architectures, and applications*, vol. 1, pp. 433–486, 1995.

[135] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

VITA

Jin Soo Cho was born in South Korea to Chong and Ok Cho. He has two younger brothers, Jin Young and Sam. After graduating from East Hamilton High School, he attended the University of Tennessee at Chattanooga. In 2015, he was rewarded the Bachelors of Science with Honors in Computer Science. In 2016, he decided to pursue a Master's degree in Computer Science: Data Science and a Ph.D. degree in Computational Science: Computer Science. Jin graduated with a Doctorate degree from the University of Tennessee at Chattanooga in May 2021.