LEARNING A LONG HEALTHY AGING

BY

FARAZ FAGHRI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Roy H. Campbell, Chair
Professor ChengXiang Zhai
Assistant Professor Jian Peng
Professor King Li, MD
Dr. Andrew B. Singleton, National Institute on Aging, National Institutes of Health
Dr. Mike A. Nalls, National Institute on Aging, National Institutes of Health

ABSTRACT

The human body is formed from the DNA code within a person's zygote. As the body is programmatically built from the zygote, the swarms of RNA/DNA greatly increase, eventually forming the neurological system and kick-starting it to become intelligent. The human body is gradually but continuously changing; we call this gradual accumulation of changes, aging. We age at different rates, in different forms, depending on many factors throughout our lifespan. Ultimately, the many physical systems that make up our body begin to fail at the same time and in mutually detrimental ways.

The human body is a machine, and like any machine, it can be modeled, predicted, and maintained for a substantial length of time. We may postpone or reduce the undesired effects of aging. Maintaining physical and mental health, avoiding disorders, and remaining active and independent. Aging is part of the human experience, and we can strive to make it positive. This forms the basis for considering wellbeing in older age, healthy aging. Quantifying the human body as a machine can illuminate what are the elements of a healthy aging process and avoid undesirable outcomes. We may predict the aging trajectory, the rate and form of changes, the occurrence of degenerative disorders such as Alzheimer's, Parkinson's, and Amyotrophic lateral sclerosis (ALS). We may prescribe lifestyle changes. We may intervene and prevent an undesirable trajectory.

In pursuit of healthy aging, we utilize data-driven approaches to learn and model the aging of the human body. We utilize the power of data analytics, machine learning, cloud computing, and well-curated datasets. We use machine learning techniques on longitudinal data to develop descriptive, predictive, and prescriptive models of aging. We focus on aging and neurological disorders as one of the most prominent health disorders in our aging population. Our solutions impact the whole spectrum of healthcare from patients and caregivers to physicians and clinicians to providers and insurers.

To my family.

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to my advisor Professor Roy H. Campbell, for his relentless support and guidance on my research. This dissertation would not be possible without his inspiring ideas, constructive criticism, and invaluable insights. Roy's deep belief in me and the freedom he gave enabled me to explore and discover new areas.

I am grateful to my mentors and great collaborators Dr. Mike Nalls, Dr. Andrew Singleton, Dr. Bryan Traynor, and Dr. Sonja Scholz for their direction and the opportunity of working on crucial health problems. I also wish to thank other members of my Ph.D. committee, Prof. ChengXiang Zhai, Prof. Jian Peng, and Prof. King Li, for their ingenious suggestions on my dissertation. I also like to thank Prof. Saurabh Sinha, Prof. Paris Smaragdis, Prof. Klara Nahrstedt, Prof. Indranil Gupta, Prof. Ravishankar Iyer, Prof. Gene Robinson, and Prof. Jiawei Han, for their help and guidance during my studies.

My sincere gratitude to my colleagues and friends at the University of Illinois at Urbana-Champaign including Sayed Hadi Hashemi, Mohammad Babaeizadeh, Adel Ahmadyan, Shadi Abdollahian Noghabi, Fardin Abdi, Read Sprabery, Imani Palmer, Sahand Mozaffari, Amin Ansari, Hassan Eslami, Amin Sadeghi, Babak Behzad, Ankit Singla, Rakesh Gopchandani, Sangeetha Abdu Jyothi, Amirhossein Taghvaei, Seyed Rasoul Etesami, Elyas Goli, Mohammad HamediRad, Biplab Deka, Reza Farivar, Behrouz Touri, Abhishek Verma, Mirko Montanari, Riccardo Crepaldi, Cristina Abad, and Shivaram Venkataraman. I'm very happy to have had the opportunity to collaborate with an amazing set of friends and researchers through my time at the National Institutes of Health (NIH), including Cornelis Blauwendraat, Sara BandresCiga, Hampton Leonard, Ruth Chia, Hirotaka Iwaki, Lana Sargent, and Aude Nicolas. I am grateful for their help, guidance, and support. I would also like to thank all of the undergraduate and graduate students that I worked with, including Vipul Satone, Rachneet Kaur, Yu-Wei Lin, Yuqian Zhou, and Fabian Brunn for their enthusiasm and hard work.

I am grateful to have had the Mayo Clinic fellowship opportunity through the Mayo Clinic and University of Illinois Alliance for Technology-Based Healthcare. I was fortunate to work with world-class physicians and scientists at the Mayo Clinic. Many thanks to Prof. Neal Cohen and Prof. Bryan White in the alliance; Dr. Bradley Erickson and Dr. Panagiotis

Korfiatis in Radiology; Dr. Paul Friedman, Dr. Peter Noseworthy, and Zachi Attia in Cardiology; Dr. Ali Daneshmand in Neurology and ICU; Steve Demuth and Nathan Spillers in Information Technology; Dr. Daniel Quest and Dr. Steven Hart in Bioinformatics and Data Science.

Many thanks to the staff of the Computer Science department, including but not limited to: Kathy Runck, Kara MacGregor, Mary Beth Kelley, Viveka Perera Kudaligama, and Maggie Metzger Chappell for their prompt help and support at multiple junctures of my studies.

But most of all, I could not have completed this journey without the support of my family. This thesis is dedicated to my parents, Laleh Kordavani and Hassan Faghri, for all the years of their love and support. My parents' endless effort has paved the way for my success. They are the reason I am where I am today. And to my brother, Fartash Faghri, for the heated debates and always challenging me. I also like to thank my uncle Ali Faghri for his tremendous encouragement, guidance, and support.

This work was supported by funding from Google and Microsoft for cloud computing services.

I thank the participants, patients and their families who contributed to this research.

This work was supported in part by the Intramural Research Program of the National Institute on Aging and the National Institute of Neurological Disorders and Stroke, National Institutes of Health, Department of Health and Human Services (project _ZO1 AG000949, ZIA-NS003154) and the Michael J Fox Foundation. This work has also been supported in part through the grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH big data to Knowledge (BD2K) initiative (`www.bd2k.nih.gov`).

Data used in the preparation of this work were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (`www.ppmi-info.org/data`). For up-to-date information on the study, visit `www.ppmi-info.org`. PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Avid Radiopharmaceuticals, Biogen Idec, Bristol-Myers Squibb, Covance, Eli Lilly & Co., F. Hoffman-La Roche, Ltd., GE Healthcare, Genentech, GlaxoSmithKline, Lundbeck, Merck, MesoScale, Piramal, Pfizer, and UCB. Data and biospecimens used in prepa-

ration of this manuscript were obtained from the Parkinson's Disease Biomarkers Program (PDBP) Consortium, part of the National Institute of Neurological Disorders and Stroke at the National Institutes of Health. Investigators include: Roger Albin, Roy Alcalay, Alberto Ascherio, DuBois Bowman, Alice Chen-Plotkin, Ted Dawson, Richard Dewey, Dwight German, Xuemei Huang, Rachel Saunders-Pullman, Liana Rosenthal, Clemens Scherzer, David Vaillancourt, Vladislav Petyuk, Andy West and Jing Zhang. The PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

In this thesis, *"Learning A Long Healthy Aging"*, our mission is to use *Machine Learning* to describe, predict, and prescribe *a Long*-itudinal *Healthy Aging*. To do so, we take a multi-disciplinary approach to the problem, which can be described by three principal components: (i) healthy aging, (ii) longitudinal and multi-modal data, (iii) machine learning and cloud computing. In this chapter, we start by providing a background on aging and human modeling. Then, we outline our approach, strategy, and keystone challenges to overcome. Finally, at the end of this chapter, we lay out the roadmap of this dissertation.

1.1 BACKGROUND

The human body is formed from the DNA code within a person's zygote. As the body is programmatically built from the zygote, the swarms of RNA/DNA greatly increase, eventually forming the neurological system and kick-starting it to become intelligent. The human body is gradually but continuously changing; we call this gradual accumulation of changes, aging. We age at different rates, in different forms, depending on many factors throughout our lifespan. Ultimately, the many physical systems that make up our body begin to fail at the same time and in mutually detrimental ways.

The human body is a machine, and like any machine, it can be modeled, predicted, and maintained for a substantial length of time. We may postpone or reduce the undesired effects of aging. Maintaining physical and mental health, avoiding disorders, and remaining active and independent. Aging is part of the human experience, and we can strive to make it positive. This forms the basis for considering wellbeing in older age, healthy aging. Quantifying the human body as a machine can illuminate what are the elements of a healthy aging process and avoid undesirable outcomes. We may predict the aging trajectory, the rate and form of changes, the occurrence of degenerative disorders such as Alzheimer's, Parkinson's, and Amyotrophic lateral sclerosis (ALS). We may prescribe lifestyle changes. We may intervene and prevent an undesirable trajectory.

In pursuit of healthy aging, we utilize data-driven approaches to learn and model the aging of the human body. We utilize the power of data analytics, machine learning, high-performance computing, and well-curated datasets. We use machine learning techniques on longitudinal data to develop descriptive, predictive, and prescriptive models of aging. We

focus on aging and neurological disorders as one of the most prominent health disorders in our aging population. Our solutions impact the whole spectrum of healthcare from patients and caregivers to physicians and clinicians to providers and insurers.

1.1.1 Health and Aging Human Body

The human body goes through physiological changes leading to senescence, the decline of biological functions. Calling this process "aging," it begins as soon as adulthood is reached. With age, bones shrink in size and density, more susceptible to fracture. Muscles lose strength, endurance, and flexibility – factors that affect coordination, stability, and balance. The brain undergoes changes affecting memory or thinking skills, hearing may diminish, vision may cloud, gums may recede, skin thins and becomes less elastic.

However, we may postpone or reduce the undesired effects of aging. Maintaining physical and mental health, avoiding disorders, and remaining active and independent. This form of aging is referred to as "healthy aging" or more formally by the World Health Organization (WHO) *"as the process of developing and maintaining the* ***functional ability*** *that enables* ***wellbeing*** *in older age"* [1].

1.1.2 Human Body as a Machine

For over a thousand years, since the time of ancient Greeks and Aristotle, the idea of *élan vital* or "vital force" was endured by many in the West. The idea that living organisms are fundamentally different from non-living entities because they contain some non-physical element or are governed by different principles than are inanimate things [2]. Until Leonardo da Vinci and Andreas Vesalius began dissections and circulated anatomical drawings in the early 1500s [3]. It became evident that bones and muscles were just systems of levers, ropes, and pulleys. However, it was not until the early 1600's that the French philosopher René Descartes replaced vitalism with scientific materialism. In 1637, Descartes, the great philosopher, mathematician, and natural scientist, published one of his most important texts, namely the *Discourse on the Method of Rightly Conducting one's Reason and Seeking the Truth in the Sciences*, commonly known as the *Discourse* [4]. He proposed the concept of body as a machine, although a very complicated one [5]. His formulation of the body-soul problem has served as the starting point for most historical inquiries. The bodies of humans and brutes according to Descartes were complex machines whose many actions and physiological functions were caused by the mechanical motions of their parts following "from

the mere arrangement of the machine's organs every bit as naturally as the movements of a clock or other automaton follow from the arrangement of its counter-weights and wheels" [6]. Descartes's metaphor of body as a machine became increasingly influential as the industrial revolution transformed society [7]. The idea dominated thinking in biology and medicine. Today, we call this new scientific approach to medicine, biomedicine, based on an underlying idea of the body as a machine that has been called 'the biomedical model' [8]. Biomedical models have improved our lives by encouraging a detailed analysis of the body's mechanisms at all levels, from the details of anatomy to understanding how hormones like insulin regulate chemicals like glucose [9]. We are now going down to lower levels of genes and molecules.

Over the past century, medicine has been modeled by the traditional medical model, one in which physicians cure biological disease using biomedical mechanistic reasoning [10]. In this conventional model, paradigmatic diseases are acute infectious diseases, which are generally curable, and can be understood and treated using biologic rationale: for a bacterial infection, treat with an antibiotic to halt the germ's growth or survival, and thus clear the infection. However, a more comprehensive and personalized medical model is needed to empower a new era of medicine. The new medical model is enshrined in "precision medicine," a medical model that customizes medical decisions, treatments, or practices tailored to the individual patient [11]. In one illustration of this approach, scientists have developed a genetic risk score for type 2 diabetes and coronary artery disease [12]. Evidently, a significant number of insulin resistance variants are highly associated with higher triglycerides, lower HDL cholesterol, and greater hepatic steatosis, despite reduced adiposity. Using this new model, investigators have found that the leaner individuals who carried a heavier genetic burden were more likely to develop type 2 diabetes or coronary artery disease. Thus, new precision medicine models can help better define patients commonly classified under the broad umbrella of type 2 diabetes and select them for targeted preventive or therapeutic interventions [13]. Unlike precision medicine-based models, traditional disease-specific models of a patient with type 2 diabetes and heart failure would have the physician manage these biological disturbances (in glucose homeostasis, in cardiac function) individually while preventing other diseases or complications.

1.1.3 Human Body Machine and Computing Machine

Arguably, the "Computer Age" started when Alan Turing in 1936 introduced his abstract 'computing machines' in his first major publication, 'On Computable Numbers, with an Application to the Entscheidungsproblem' [14]. Turing presented a theoretical machine, now

called 'Turing Machine,' that could solve any problem that could be described by simple instructions encoded on a paper tape. Later on, likely influenced by 300-year old Descartes's metaphor of the body as a machine, Alan Turing in his 1950 *Mind* paper 'Computing Machinery and Intelligence,' [15] introduces what is now called 'The Turing test' [16]. Turing had an idea that computers would become so powerful that they would think. He envisaged a time when artificial intelligence (AI) would be a reality.

Today, when we must deal with problems that are too complicated for our brains, we resort to mathematical models and computers. Take weather forecasting: to predict tomorrow's weather, we need to take into account so many factors, and so many calculations, that it could take months, if not years, to come up with an answer. But if we write all we know about weather in a mathematical model, a computer can do those calculations fast enough to predict the weather for tomorrow in a few hours. Similar mathematical modeling approach (a.k.a. simulation-based) has been used in medicine as early as 1979 [17] by developing a whole generation of biomedical simulation models with the aim of predicting what will happen in the human bodies under a variety of conditions. Generally, simulation-based modeling describes a physiological body as a combination of differential equations. For instance, using fluid mechanic models for the analysis of cardiovascular function [18]. This physics-based simulation modeling requires extensive physical and operational knowledge of a target system in order to be accurate. This required knowledge is scarce even for the simplest part of our body.

With the rise of "big data" in healthcare, data-driven modeling has become more feasible. There are various sources of big medical data, such as administrative claim records, clinical registries, electronic health records, biometric data, patient-reported data, the internet, medical imaging, biomarker data, genetic data, prospective cohort studies, and large clinical trials [19]. However, data has not been the only catalyst for the rise of data-driven methods in medicine. Advances in analytic techniques in computer science, especially in machine learning, has been a significant contributor [20]. In the last couple of years, breakthroughs started happening in machine learning. Techniques began working much better, while new techniques have appeared, especially around artificial neural networks, and when they were applied to some long-standing and important use cases, better results were gained [21]. These techniques have also shown to have much broader applications, especially in medicine [22, 23, 24].

1.1.4 Awareness of Limitations

"All models are wrong, but some are useful," a point well made by George Box in his often-cited remark [25, 26]. Furthermore, he reminds us that "the practical question is how wrong do they have to be to not be useful" [27]. In hindsight, the idea of the body as a machine has led to many improvements in medicine; however, one should be aware that it forges beliefs about the body that oversimplify ideas of how the body works and how the disease works [28, 29]. Because humans are evolved, they are fundamentally different from human-made machines. Recognizing the limitations, in this work, we will point out model approximations where the machine fails to recognize evolution and organic complexity. Ultimately, a body is a body is a body is a body[1], but as John von Neumann said: "truth ... is much too complicated to allow anything but approximations" [30].

1.2 THESIS STATEMENT

Machine learning enables the construction of the human aging model. Descriptive, predictive, prescriptive machine learning methods contribute to this human aging model and permit health-related decisions for clinicians and individuals.

1.3 DISSERTATION FOUNDATIONS

In this thesis, *"Learning A Long Healthy Aging"*, our mission is to use *Machine Learning* to describe, predict, and prescribe *a Long*-itudinal *Healthy Aging*. To do so, we take a multi-disciplinary approach to the problem, which can be described by three principal components: (i) healthy aging, (ii) longitudinal and multi-modal data, (iii) machine learning and cloud computing. Here we describe each component in more detail:

1.3.1 Healthy Aging

The human body goes through physiological changes leading to senescence, the decline of biological functions. Calling this process "aging," it begins as soon as adulthood is reached (Figure 1.1). Aging is part of the human experience, and we can strive to make it positive. This forms the basis for considering wellbeing in older age, healthy aging.

[1]Homage to "Rose is a rose is a rose is a rose" by Gertrude Stein as part of the 1913 poem "Sacred Emily."

In this thesis, we focus on brain aging and neurological disorders. This subset of aging-related diseases is representative of a large portion of aging trajectories. According to a recent study in the Netherlands, 1 in 2 women and 1 in 3 men set to develop dementia, parkinsonism, or stroke during their lifetimes [31]. They further show that preventive strategies that delay disease onset of all three diseases by 1-3 years have the potential to reduce these risks by 20%-50%. This study further highlights how taking proactive healthy lifestyle measures, and early interventions can significantly lessen the risk of these diseases, regardless of age. The global costs-of-illness for these diseases is estimated to amount to more than 2% of the world's annual gross domestic product (GDP), a figure that is set to rise steeply with the aging of populations and continuing increases in life-expectancy worldwide [32, 33, 34, 35]. As a result, prioritizing these diseases on the global health agenda is crucial [36].

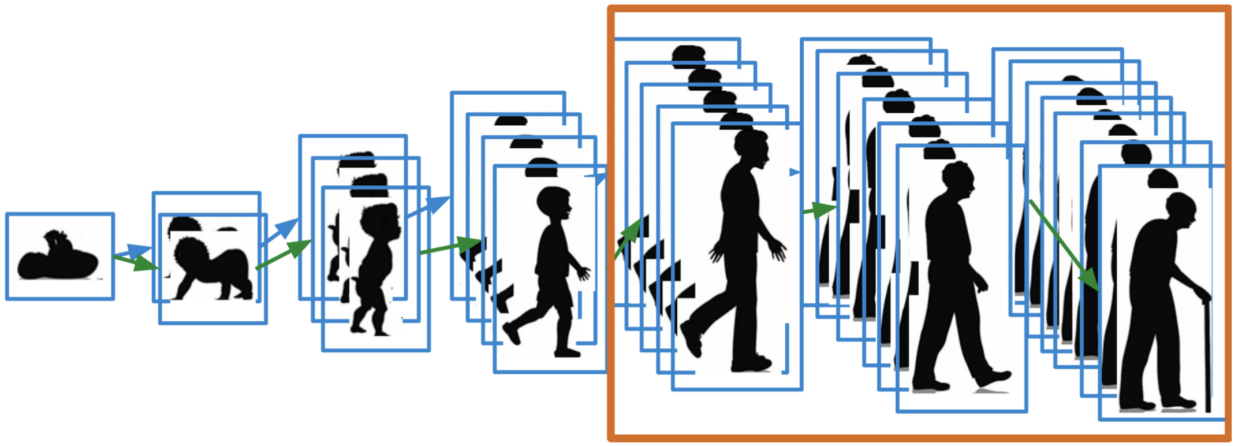

Figure 1.1: Aging of the human body and the quest for healthy aging trajectories.

1.3.2 Longitudinal and Multi-modal Data

Shown in Figure 1.2, our data-driven strategy starts with the collection and integration of longitudinal studies with highly dimensional multi-modal datasets. In order to develop a timely model of aging, we need to incorporate longitudinal studies with time-series data. This includes both short-term and long-term longitudinal data. Short-term data such as sensory and monitoring device data has a high frequency of data collection, but with a time window of hours to days. Long-term data, such as clinical assessments, have a low frequency of data collection with a time window of six months to two years.

Historically, studies biomedical studies have focused on one type of data, e.g., only genetic or imaging. However, the human body is remarkably complex, and no single modality can

capture its manifestation in full. Our approach is to utilize any available modality, including clinical, biological, imaging, environmental, lifestyle, genomics, and multi-omics. We hope by integrating all of them, we can have a stronger aging model and a better understanding of the human body.

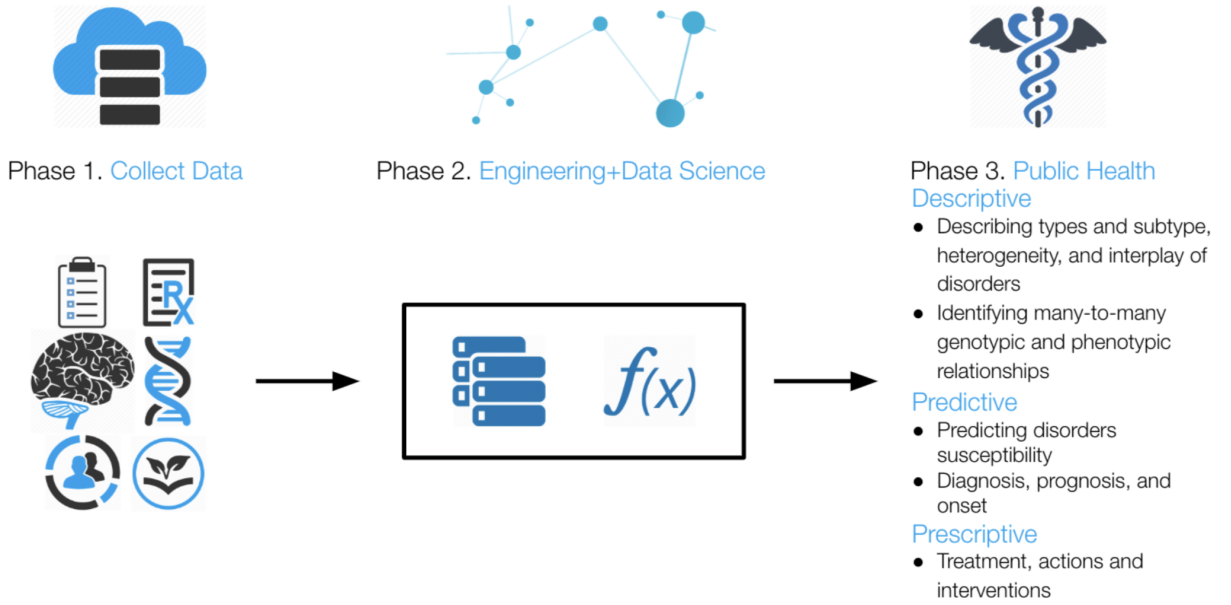

Figure 1.2: The data-driven strategy of the thesis. Collection and integration of multi-modal longitudinal data, use of machine learning and cloud computing solutions, and delivering public health benefits.

1.3.3 Machine Learning and Cloud Computing

With longitudinal and multi-modal data in hand, the next principal component is to develop the human aging model. We utilize machine learning techniques, including supervised, unsupervised, and semi-supervised learning, to analyze the data and develop models. Historically, *descriptive* modeling has been applied to health data to identify genotypic and phenotypic relationships. We are now moving toward *predictive* analytics, building on the capabilities of descriptive analytics to forecast future onset, diagnosis, and prognosis using various models and what-if analyses. In the long term, we will utilize *prescriptive* analytics to forecast possible outcomes and allow providers to make proactive decisions. We use these models to predict the body's aging trajectory that may guide us manipulate what would otherwise happen, unhealthy aging. We predict the rate and form of changes, the occurrence of degenerative disorders such as Alzheimer's, Parkinson's, and ALS. Ultimately,

we may prescribe lifestyle changes. We may intervene and prevent an undesirable trajectory.

Analyzing health data and particularly imaging and genomics data, pose severe computational challenges. Considering the computational demands across the lifecycle of a dataset –acquisition, storage, distribution, and analysis– genomics is either on par with or the most demanding of the big data domains [37]. To address the storage and computational challenges posed by big health data, we design and use cloud computing-based analytical systems that have transitioned from shared, centralized architectures to distributed, decentralized architectures. We perform large-scale computing and utilize storage resources for open science data sharing. However, in the cloud environment, the computation time is costly, to reduce the time and cost, it is essential to implement and use optimized machine learning algorithms that intelligently reduce the network overhead and I/O waste while utilizing CPU resources.

1.4 DISSERTATION OUTLINE

The contributions of this thesis are organized into chapters according to the studied disorder and clinical setting:

- In Chapter 2, we study Parkinson's disease. Using an unsupervised learning approach, we identify new subtypes of the disorder based on disease progression. We also provide an in-depth analysis of these subtypes. Furthermore, we develop predictive models for early diagnosis, prognosis, and clinical trial stratification. This work previously appeared as [38, 39, 40, 41, 42, 43, 44, 45, 46, 47].

- In Chapter 3, we present our work on Alzheimer's disease. Similar to the work on Parkinson's disease, we use an unsupervised learning approach to identify new subtypes of the disorder based on disease progression. We also provide an in-depth analysis of these subtypes. Furthermore, we develop predictive models for early diagnosis, prognosis, and stratification. This work previously appeared as [48, 49, 50, 51].

- In Chapter 4, we introduce our work on Amyotrophic lateral sclerosis (ALS) disorder. Similar to Alzheimer's and Parkinson's chapters, we use an unsupervised learning approach to identify new subtypes. However, unlike Parkinson's and Alzheimer's, which have a long multi-year course, we do not predict progression. Since ALS is a rapidly progressive disorder, we focus on subtype identification and survival analysis. We also use semi-supervised learning for enhancing subtype identification. A subset of this work was published previously as [52, 53, 54].

- In Chapter 5, we present the work on readmission prediction in the Intensive Care Unit (ICU). Unlike previous chapters where the health issue has a multi-year course, in the ICU, data is short-term. Short-term data such as sensory and monitoring device data has a high frequency of data collection, but with a time window of hours to days. Due to the data collection frequency, different predictive machine learning methods are used to encapsulate the time-sensitivity of small fluctuations in the patient's status. This work was published previously as [55].

- Finally, we conclude by discussing lessons learned and open challenges in learning a long healthy aging in Chapter 6.

Some of the work that embodies various topics in computer science related to this thesis, but not described in the thesis, includes the work on cloud computing reliability [56], big data, and genomic [37], scalable machine learning [57], privacy-preserving data distribution [58, 59], and scalable genotyping [60].

1.4.1 Keystone Challenges and Contributions

In this dissertation and in the endeavor of learning the aging model, we have faced two categories of challenges: engineering and clinical. This work has contributed to tackling both categories. The models in this work are designed for different scenarios, but share the following designing principles and address four engineering challenges:

1. *Lack of data for supervised learning.* In order to develop accurate predictive models based on supervised learning, we need large and reliable data. In healthcare, such data is mostly not available; labels come from physicians who themselves have a high misdiagnosis rate. To overcome this challenge, this work heavily relies on labeling data using *unsupervised* and *semi-supervised learning* techniques.

2. *Utilization of short-term and long-term longitudinal data.* In order to develop a timely model of aging, we need to incorporate longitudinal studies with time-series data. In healthcare, we have both *short-term* and *long-term* longitudinal data. Short-term data such as sensory and monitoring device data has a high frequency of data collection, but with a time window of hours to days. On the other hand, long-term data such as clinical assessments have a low frequency of data collection with a time window of six months to two years. In this work, we utilize both types of longitudinal data.

3. *Integration of multi-modal data.* Many machine learning techniques focus on one type of data, e.g., only imaging or audio. However, in healthcare, we look at the human

body from different modalities, and we hope by integrating all of them, we can have a better understanding of the issue. Utilizing highly dimensional multi-modal datasets, including clinical, biological, genetic, and imaging data have been part of this work.

4. *Interpretability.* Using unsupervised machine learning techniques, we have developed *embedding spaces* that have guided us in labeling the subjects. Part of this labeling relies on our success in interpreting the embedding spaces. In this work, we have put effort into dissecting the machine learning "black box" to understand the results better and guide the development of models.

We also address the following clinical challenges:

1. *Replication of results with other datasets.* Validating findings and replication is an underpinning of research. Generalizing, the same methods and protocols should be used on a different group of people, or a different setting, and come up with similar results. In healthcare, lack of data has made replication more sparse. However, in this work, when possible, we show that results are valid in external datasets.

2. *Developing usable models in both clinical and research settings.* Models with higher accuracy demonstrate a more powerful screening capability in assisting the physicians. However, accuracy is not enough, and we need to further evaluate the machine learning models for use in a clinical setting. We need to assess the accuracy along with operating points corresponding to sensitivity (also called the true positive rate, the recall, or probability of detection) and specificity (true negative rate) of the algorithm with respect to the reference standards [61]. It is often claimed that these targeted operating points can be used for different clinical purposes; for instance, a highly sensitive test is deemed effective at ruling out a disease when negative, whereas a highly specific test is effective at ruling in a disease when positive. However, these rules are misleading, as the diagnostic power of any test is determined by both its sensitivity and its specificity [62]. The tradeoff between specificity and sensitivity is explored in ROC analysis [63]. In this work, we analyze the usability of all the models for clinical settings, not only the individual operation points but across a range of values for the ability to predict a dichotomous outcome.

3. *Tangible improvement to physician's decision making.* Feature interpretation, as well as decision making logic, reliability, and robustness analysis of the machine learning models, is crucial and imperative for clinical applications. This task is much more complicated for recent techniques. Many recent efforts are short of explaining the

decision-making logic and model interpretation in healthcare. In this work, we dive deeper into our machine learning model in an effort to further interpret the results, capabilities, and limitations. We investigate the most important factors that the machine learning model has learned in order to predict and classify an event. We review the clinical literature for additional verification and a better clinical understanding of the machine learning model. Finally, we examine the advantages and strengths of the proposed models.

CHAPTER 2: LEARNING PARKINSONISM AND PARKINSON'S DISEASE

In this chapter, we review the work on *Predicting onset, progression, and clinical subtypes of Parkinson's disease using machine learning*. Using an unsupervised learning approach, we identify new subtypes of the disorder based on disease progression. We also provide an in-depth analysis of these subtypes. Furthermore, we develop predictive models for early diagnosis, prognosis, and clinical trial stratification. This work previously appeared as [38, 39, 40, 41, 42, 43, 44, 45, 46, 47].

The clinical manifestations of Parkinson's disease are characterized by heterogeneity in age at onset, disease duration, rate of progression, and a constellation of motor versus non-motor features. Due to these variable presentations, counseling of patients about their individual risks and prognosis is limited. There is an unmet need for predictive tests that facilitate early detection and characterization of distinct disease subtypes as well as improved, individualized predictions of the disease course. The emergence of machine learning to detect hidden patterns in complex, multi-dimensional datasets provides unparalleled opportunities to address this critical need.

We used unsupervised and supervised machine learning approaches for subtype identification and prediction. We used machine learning methods on comprehensive, longitudinal clinical data from the Parkinson's Disease Progression Marker Initiative (PPMI) (n=328 cases) to identify patient subtypes and to predict disease progression. The resulting models were validated in an independent, clinically well-characterized cohort from the Parkinson's Disease Biomarker Program (PDBP) (n=112 cases). Our analysis distinguished three distinct disease subtypes with highly predictable progression rates, corresponding to slow, moderate, and fast disease progressors. We achieved highly accurate projections of disease progression four years after initial diagnosis with an average Area Under the Curve of 0.93 (95% CI: 0.96 ± 0.01 for PDvec1, 0.87 ± 0.03 for PDvec2, and 0.96 ± 0.02 for PDvec3). We have demonstrated robust replication of these findings in the independent validation cohort.

These data-driven results enable clinicians to deconstruct the heterogeneity within their patient cohorts. This knowledge could have immediate implications for clinical trials by improving the detection of significant clinical outcomes that might have been masked by cohort heterogeneity. We anticipate these machine learning models will improve patient counseling, clinical trial design, allocation of healthcare resources, and ultimately individualized clinical

care.

2.1 INTRODUCTION

Parkinson's disease (PD) is a complex, age-related neurodegenerative disease that is defined by a combination of core diagnostic features, including bradykinesia, rigidity, tremor, and postural instability [64, 65]. Substantial phenotypic heterogeneity is well recognized within the disease, which complicates the design and interpretation of clinical trials and limits counseling of patients about their disease risk and prognosis. The clinical manifestations of PD vary by age at onset, rate of progression, associated treatment complications, as well as the occurrence and constellation of motor/nonmotor features.

The phenotypic heterogeneity that exists within the PD population poses a major challenge for clinical care and clinical trial design. A clinical trial has to be suitably powered to account for interindividual variability, and as a consequence, they are either large, long, and expensive, or only powered to see dramatic effects. This problem becomes particularly burdensome as we move increasingly toward early-stage trials when therapeutic interventions are likely to be most effective. The ability to predict and account for even a proportion of the disease course has the potential to reduce the cost of clinical trials significantly and to increase the ability of such trials to detect treatment effects.

Attempts thus far at the characterization of disease subtypes have followed a path of clinical observation based on age at onset or categorization based on the most observable features [66]. Thus, the disease is often separated into early-onset versus late-onset disease, slow progressive "benign" versus fast progressing "malignant" subtypes, PD with or without dementia, or based on prominent clinical signs into a tremor-dominant versus a postural instability with gait disorder subtype [67, 68]. This dichotomous separation, while intuitive, does not faithfully represent the clinical features of the disease, which are quantitative, complex, and interrelated. A more realistic representation of the disease and disease course requires a transition to a data-driven, multi-dimensional schema that encapsulates the constellation of interrelated features and affords the ability to track (and ultimately predict) change.

Previous studies used cluster analysis, a data-driven approach, to define two to three clinical Parkinson's disease subtypes [69, 70, 71]. Depth of phenotypic information, as well as longitudinal assessments in these studies, was variable and often limited to certain clinical

features and short-term follow-ups. Moreover, many previous studies were limited by insufficient methods to capture longitudinal changes over multiple assessment visits. A recent study used cluster analysis to identify patient subtypes and their corresponding progression rates [71]. However, this study evaluated clusters according to only two time points, baseline, and short-term follow-up, that were aggregated into a Global Composite Outcome score. In return, the subtypes did not capture the fluctuations in the prognosis of subtypes. Finally, in order to be used in practice, subtyping solutions need to be replicated in a different cohort and to show the reliability of methods in assigning individual patients to a subtype.

We have previously used multi-modal data to produce a highly accurate disease status classification and to distinguish PD-mimic syndromes from PD [72]. This effort demonstrated the utility of data-driven approaches in the dissection of complex traits and has also led us to the next logical step in disease prediction: augmenting a prediction of whether a person has or will have PD to include a prediction of the timing and direction of the course of their disease.

Thus, here, we describe our work on the delineation and prediction of the clinical velocity of PD. The first stage of this effort requires the creation of a multi-dimensional space that captures both the features of the disease and the progression rate of these features (i.e., velocity). Rather than creating a space based on *a priori* concepts of differential symptoms, we used data dimensionality reduction methods on the complex clinical features observed at 48 months after initial diagnosis to create a meaningful spatial representation of each patient's status at this time point. After creating this space, we used unsupervised clustering to determine whether there were clear subtypes of disease within this space. This effort delineated three distinct clinical subtypes corresponding to three groups of patients progressing at varying velocities (i.e., slow, moderate, and fast progressors). These subtypes were validated and independently replicated. Following the successful creation of disease subtypes within a progression space, we created a baseline predictor that accurately predicted an individual patient's clinical group membership four years later. This highlights the utility of machine learning as ancillary diagnostic tools to identify disease subtypes and to project individualized progression rates based on model predictions.

2.2 METHODS

2.2.1 Study design and participants

This study included clinical data from the following cohorts: the Parkinson Progression Marker Initiative (PPMI, `http://www.ppmi-info.org/`; n=328 PD cases including 114 (35%) female; 172 controls including 66 (38%) female), and the Parkinson Disease Biomarkers Program (PDBP, `https://pdbp.ninds.nih.gov/`; n=112 PD cases including 53 (47%) female; 45 controls including 25 (56%) female). PPMI average age of PD cases was 67±9.8 and control 66.2±11.1. PDBP average age of PD cases 65±8.6 and control 63.7±9.1. The PPMI and PDBP cohorts consist of observational data from comprehensively characterized PD patients and matched controls. All PD patients fulfilled the UK Brain Bank Criteria. Control subjects had no clinical signs suggestive of parkinsonism, no evidence of cognitive impairment, and no first-degree relative diagnosed with PD. Both cohort's data went through triage for missing data, 48-month assessment, and comprehensive phenotype collection. Age and MDS-UPDRS Part III (objective motor symptom examination by a trained neurologist) distribution of cohorts at baseline were investigated using Kernel Density Estimation (KDE) to show these independent cohorts are identically distributed and ensure the integrity of replication and validation.

Each contributing study abided by the ethics guidelines set out by their institutional review boards, and all participants gave informed consent for inclusion in both their initial cohorts and subsequent studies.

For each cohort, a comprehensive and shared set of longitudinally collected common data elements were selected for analysis. We used the following data:

i International Parkinson's disease and Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part I, Part II, and Part III [73]

ii Cranial Nerve Examination

iii Montreal Cognitive Assessment [74]

iv Hopkins Verbal Learning Test [75]

v Semantic Fluency test [76]

vi WAIS-III Letter-Number Sequencing Test [77]

vii Judgment of Line Orientation Test [78]

viii Symbol Digit Modalities Test [79]

ix SCOPA-AUT [80]

x State-Trait Anxiety Inventory for Adults [81]

xi Geriatric Depression Scale [82]

xii Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease [83]

xiii REM-Sleep Behavior Disorder Screening Questionnaire [84]

xiv Epworth Sleepiness Scale [85]

2.2.2 Procedures and statistical analysis

To accompany this report, and to allow independent replication and extension of our work, we have made the code publicly available for use by non-profit academic researchers (`https://github.com/ffaghri1/PD-progression-ML`). The code is part of the supplemental information; it includes the rendered Python Jupyter notebook with full step-by-step statistical and machine learning analysis. For readability, machine learning parameters have been described in the Python Jupyter notebook and not in the text of the paper. Figure 2.1 illustrates a summary of our analysis workflow. As a first step, we transformed the dataset into a mathematically meaningful and naturally interpretable format. To achieve this objective, we a) *normalized* and b) *vectorized* all longitudinal data. Specifically, we first vectorized by transforming all observations of a particular parameter in a column vector, then appended all parameters together. We then used the *min-max* method to normalize the data. The min-max method is preferred for multi-modal longitudinal datasets compared to z-score, as it preserves the progression pattern.

To develop an interpretable representation of high modality longitudinal data, we next used the dimensionality reduction techniques. Dimensionality reduction techniques helped us to build the "progression space" where we can approximate each patient's position after the 48-month period. We used the Non-negative Matrix Factorization (NMF) technique to achieve this aim [86, 87]. Alternative methods, such as principal component analysis and independent component analysis, did not perform as well as NMF on longitudinal clinical

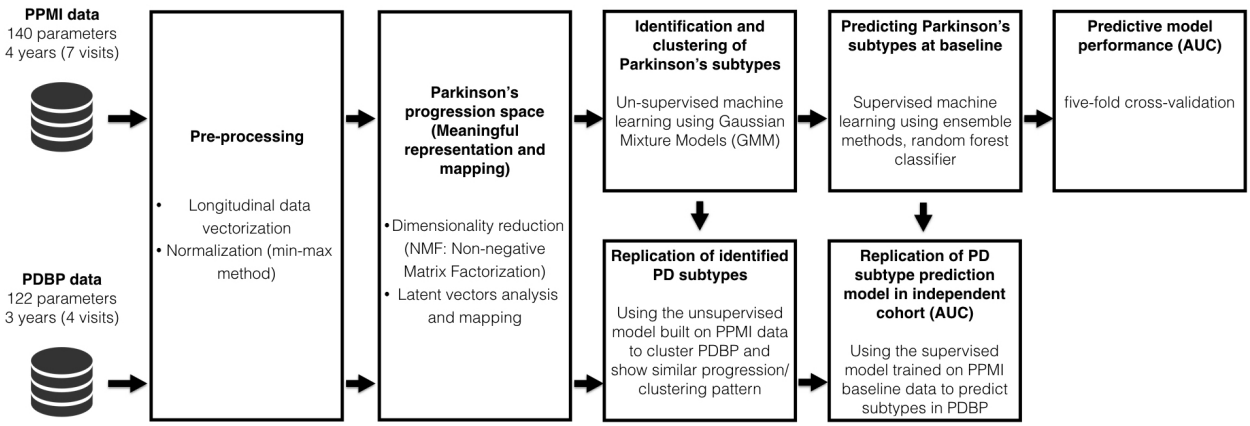

Figure 2.1: Workflow of analysis and model development.

data. This was expected due to the non-negative nature of our clinical tests. This process essentially collapses mathematically related parameters into the same multi-dimensional space; mapping similar data points close together.

Mathematically, NMF factorizes (deconstructs) the data into two matrices. Given a non-negative matrix $X \in \mathbb{R}^{m \times n}$, a non-negative decomposition of the matrix X is a pair of non-negative matrices $U \in \mathbb{R}^{m \times p}$ and $V \in \mathbb{R}^{p \times n}$ such that $X = UV$. A large number of patient parameters are aggregated in a model that represents the underlying progression concept. In this particular use case of NMF, the matrix U contains the progression space latent vectors, and the matrix V contains progression stand indicators corresponding to the latent vectors. Latent variables link observation data in the real world to symbolic data in the modeled world. By further looking into the matrix with progression space's latent vectors, we can identify the mapping and, consequently, the implications (symbolic dimensions of the modeled progression space).

Through our use of NMF, we identify primary progressive symptomatologies of the motor, cognitive, and sleep-based disturbances. Following this, unsupervised learning via Gaussian Mixture Models (GMM) [88] allowed the data to naturally self-organize into different groups relating to velocity of decline across symptomatologies, from non-PD controls representing normal aging to PD subtypes. GMM is a variant of mixture models, compared to other methods, the parametrization of a GMM allows it to efficiently capture products of variations in natural phenomena where the data is assumed generated from an independent and identically distributed mixture of gaussian (normal) distributions. The assumption of normal distribution (and therefore the use of GMM) is often used for population-based cohort

phenomenon [89].

We use the Bayesian Information Criterion (BIC) to select the number of PD clusters (subtypes) [90]. The BIC method recovers the true number of components in the asymptotic regime (i.e., much data is available, and we assume that the data was generated i.i.d. from a mixture of Gaussian distributions). To replicate the subtype identification, we apply the GMM model developed in the PPMI data to an independent cohort with varying recruitment strategy and design: the PDBP cohort. We show that identified subtypes in the PDBP are similar to the ones in the PPMI in terms of progression.

After identifying progression classes using unsupervised learning, we built predictive models utilizing supervised machine learning via ensemble methods, random forest classifier [91]. In preliminary testing, this approach outperformed other methods, such as support vector machines (SVM) and simple lasso-regression models. Besides the predictive performance, we chose random forest (RF) due to the nature of our data and problem: (i) RF is intrinsically suited for multi-class problems, while SVM is intrinsically two-class, (ii) RF works well with a mixture of numerical, categorical, and various scale features, (iii) RF can be used to rank the importance of variables in a classification problem in a natural way which helps the interpretation of clinical results, (iv) RF gives us the probability of belonging to a class, which is very helpful when dealing with individual subject progression prediction. We develop three predictive models to predict the patient's progression class after 48 months based on varying input factors: (a) from baseline factors and (b) from baseline and first-year factors. We also use a feature extraction method, Recursive Feature Elimination (RFE), in order to find significant parameters in our models.

We validate the effectiveness of our predictive models in two ways. First, we validate the algorithm using five-fold cross-validation. We randomly divided the PPMI dataset into five subsamples, retained a single subsample as the validation data for testing the model, and the remaining four samples used as training data. We repeated the process five times (the folds), with each of the subsamples used exactly once as the validation data. The performance of the algorithm in each fold was measured by the area under the receiver operating curve (AUC) generated by plotting sensitivity vs. $(1-\text{specificity})$. The five results from the folds were averaged to produce a single estimation of performance.

To conclusively validate the algorithm, we also evaluated the performance of the predictive models (trained on the PPMI measurements) on the independent PDBP cohort. We show

that the predictive models preserve their high accuracy applied to another dataset.

2.3 RESULTS

Figure 2.2 shows the result of the mathematical projection of PD progression, called *Parkinson's disease progression space*. This space shows the relative progression velocity of each patient in 48 months (i.e., speed and direction). The level of progression velocity is broken down into three main dimensions: motor, cognitive, and sleep-related disturbances. Across these trajectories, the unsupervised learning analysis reveals and classifies patients into three main subtypes of PD, relating to rates of disease progression: PDvec1, PDvec2, and PDvec3. This shows that we can now map the primary clinical symptomatology and disease progression velocity from diagnosis in Parkinson's.

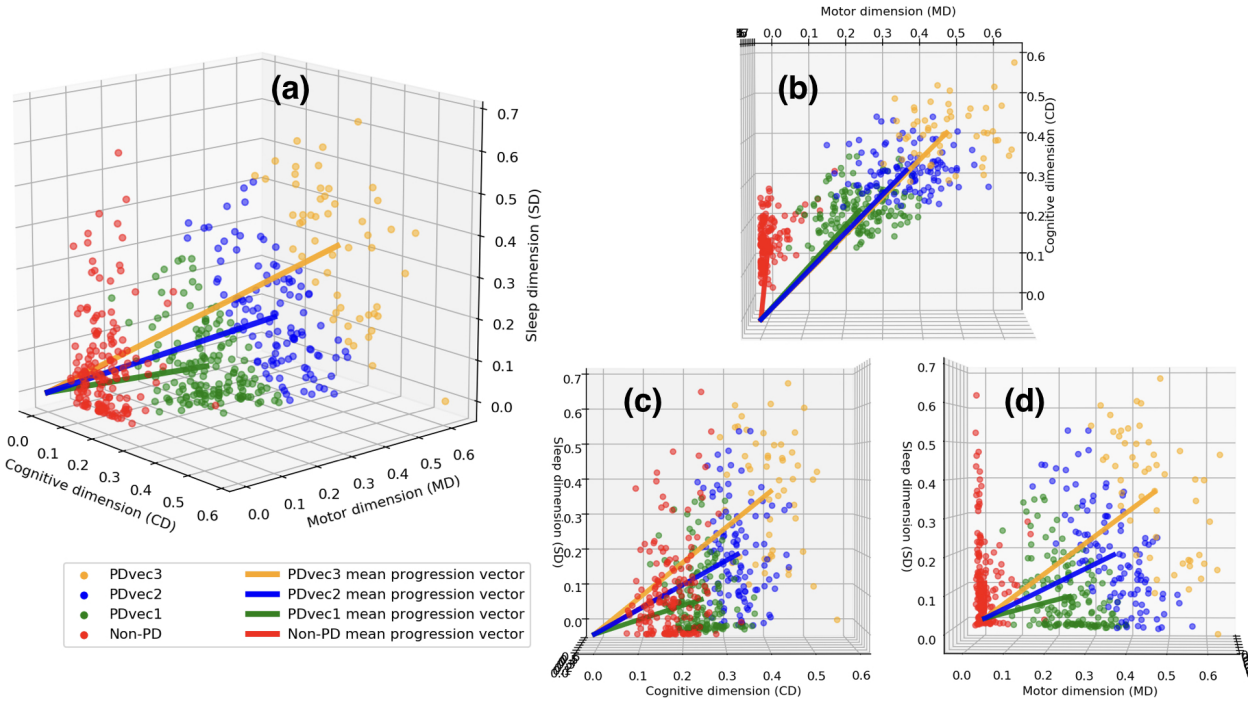

Figure 2.2: Different views of the Parkinson disease progression space with three corresponding projected dimensions (cognitive, motor, and sleep dimensions). Subtypes of PD are identified using unsupervised learning (PD vector 1, PD vec 2, and PD vec 3). (a) shows the view of all three dimensions, (b) view of the motor and cognitive dimensions, (c) view of cognitive and sleep dimensions, and (d) view of sleep and motor dimensions.

Regarding the interpretation of PD progressions space dimensions, Figure 2.3 shows the mapping guide for how the PPMI's high-dimensional space of 140 different clinical parame-

ters is mapped to the three-dimensional embedding of Parkinson's disease progression space. The columns represent the projected three dimensions, i.e., motor, cognitive, and sleep-related trajectories, and the rows are the PPMI clinical parameters. This figure allows us to not only observe the conversion but also the heterogeneity of some clinical parameters, for instance, how some of the Epworth Sleepiness Scale parameters reflect both sleep and cognitive disorders, and some reflect both sleep and movement disorders. In comparisons of the eigenvalues within the NMF decomposition, the projected motor dimension was responsible for 41.6% of the explained variance, followed by the sleep dimension (29.5%), and cognitive dimension (28.9%).

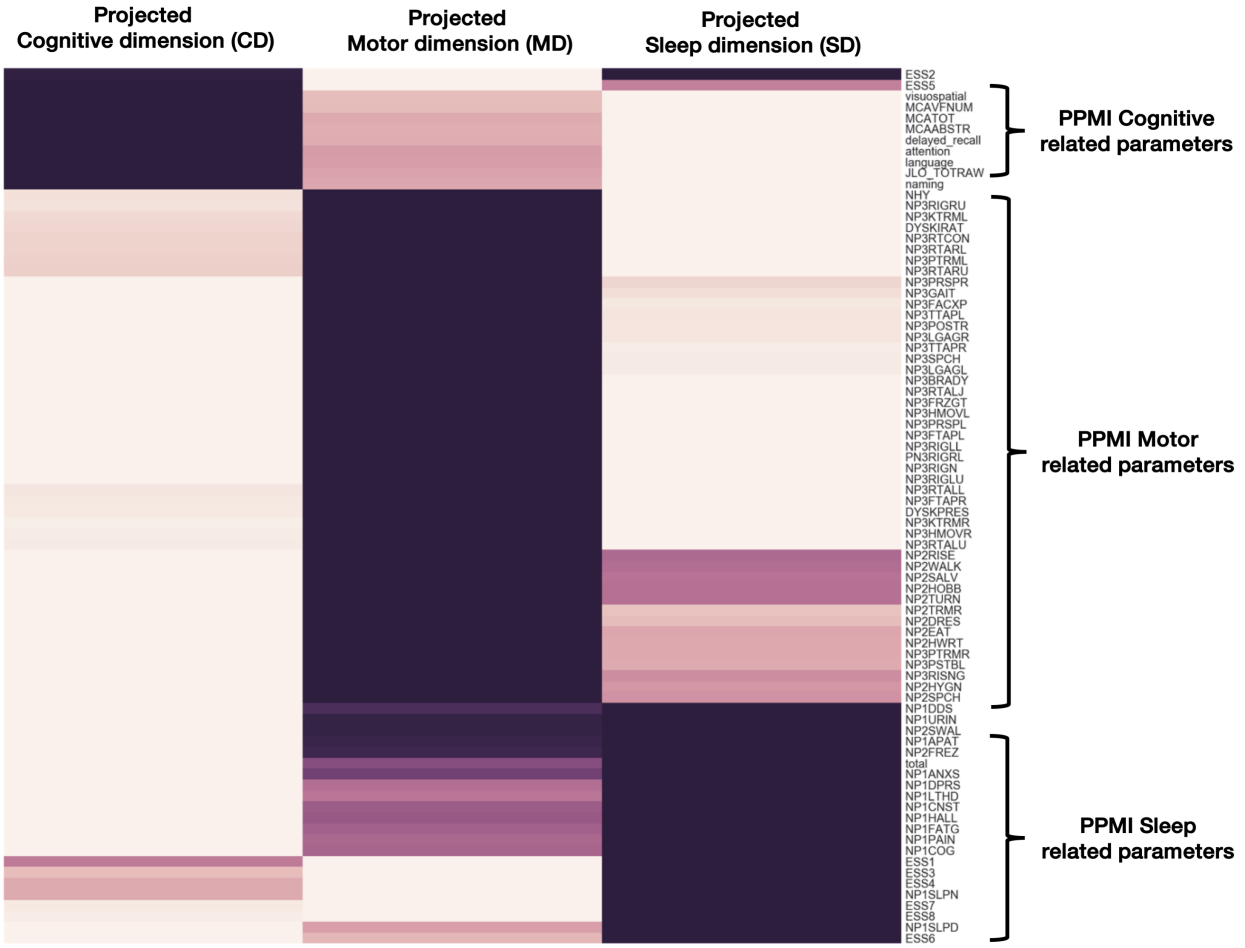

Figure 2.3: Shows how each 140 different input parameters have been projected into the new dimension of the Parkinson's progression space (cognitive, motor, and sleep dimensions). Darker colors represent strong mapping.

Regarding the number of identified PD subtypes, Figure 2.4 shows the characteristics of PDs identified subtypes in more detail. Specifically, Figure 2.4 shows the visualization of un-

supervised learning via GMM. GMM fits the data into different subtypes relating to velocity of decline across symptomatologies, from non-PD controls. The BIC method has identified three Gaussian distributions representing three PD subtypes.

In terms of characteristics of PDs identified subtypes, Figure 2.5 demonstrated how cognitive, motor and sleep-related symptoms differ within each PDs subtype and in controls. There is a clear trend for increased sleep and motor disturbances after four years in fast progressors compared to the slower progressing subtypes, which seem to have relatively more cognitive issues early on.

Figure 2.6 shows the progression of each PD subtype overtime at baseline and after 18 months, 24, 36, and 48 months. To better understand the clinical presentation of the three identified subtypes, Figure 2.6 demonstrates the three main projected dimensions (motor, cognitive and sleep-related disturbances), as well as actual clinical values of each subtype overtime for UPDRS-Part I, Part II, Part III, as well as Hopkins Verbal Learning Test, Symbol Digit Modalities Test, Semantic Fluency test, Epworth Sleepiness Scale, State-Trait Anxiety Inventory for Adults, and Geriatric Depression Scale.

In terms of the genetic association of PDs identified subtypes, Genetic Risk Scores (GRS) were calculated as per [46]. While the GRS was not selected during feature extraction in the clustering exercise we did analyze regressions comparing associations between the GRS and either the continuous predicted cluster membership probability (linear regression) or the binary membership in a particular cluster group compared to the others. All models were adjusted for age at onset, female gender, and principal components as covariates to adjust for population substructure in PPMI. The GRS was significantly associated with decreasing magnitude of the sleep vector per Standard deviation of increase in the GRS ($beta$ = -0.0298546, se = 0.0097124, p = 0.00232, adjusted r^2 = 0.04584). For binary models of membership, we see that the GRS is weakly but significantly associated with a decreased risk of membership in PDvec3 (odds ratio = 0.5630876 per 1 SD increase from case GRS mean, $beta$ = -0.574320, se = 0.243974, P = 0.01857) and increased risk of membership in PDvec1 (odds ratio = 1.340952, $beta$ = 0.29338, se = 0.13367, P = 0.028178). The lack of a strong genetic association is due to the small sample size.

In order to ensure the generalizability and validity of the results, we replicate the subtype identification in the independent PDBP cohort. Figure 2.7 shows PPMI and PDBP cohorts are similarly distributed; hence, they are suitable for replication and validation. Further-

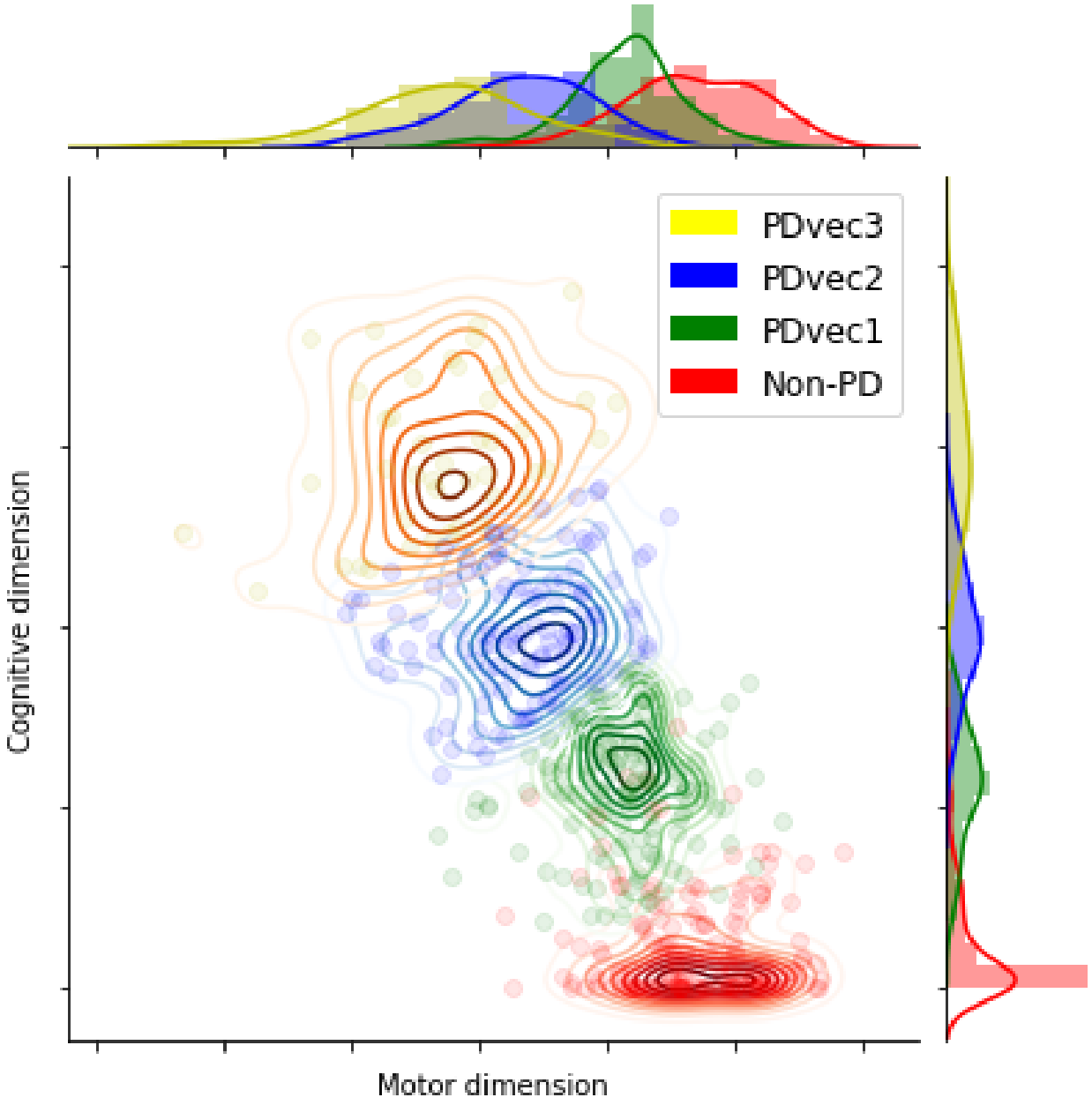

Figure 2.4: Visualization of unsupervised learning via GMM and identification of three Gaussian distributions representing three distinct PD subtypes. Motor dimension reflects the increase in disturbance, while the cognitive dimension reflects the decline.

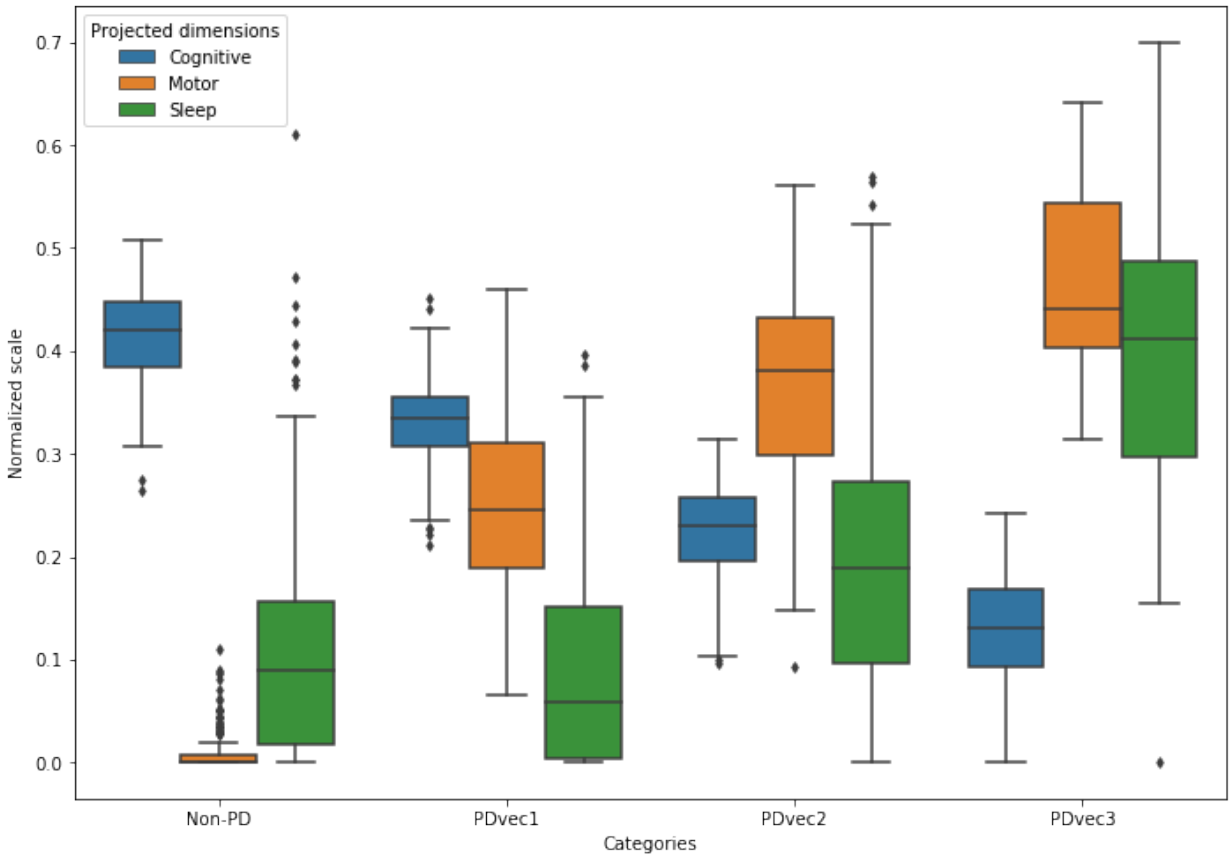

Figure 2.5: Shows the distribution of projected dimensions (cognitive, motor, and sleep) for each Parkinson's category and healthy control after three years. Motor and sleep dimensions reflect the increase in disturbance, while the cognitive dimension reflects the decline. PDvec3 has the highest motor and sleep disturbance, as well as the highest cognitive decline.

more, we have performed the two-sample t-test for quantified replication cohort validation analysis (Table 2.1).

Figure 2.8 shows the identified subtypes in the independent PDBP cohort using the model developed on the PPMI dataset. We see that the identified subtypes in the PDBP are similar to the ones in the PPMI in terms of progression. The PPMI and PDPB cohorts are clinically different cohorts and recruited from different populations. The replication of our results in the PDBP cohort that was recruited with a different protocol shows the strength of our study's methodology. We demonstrate that if we ascertain the same phenotypes using standardized scales, we can reliably discern the same subtypes and progression rates. This makes our results generalizable and the clinical subtypes reproducible.

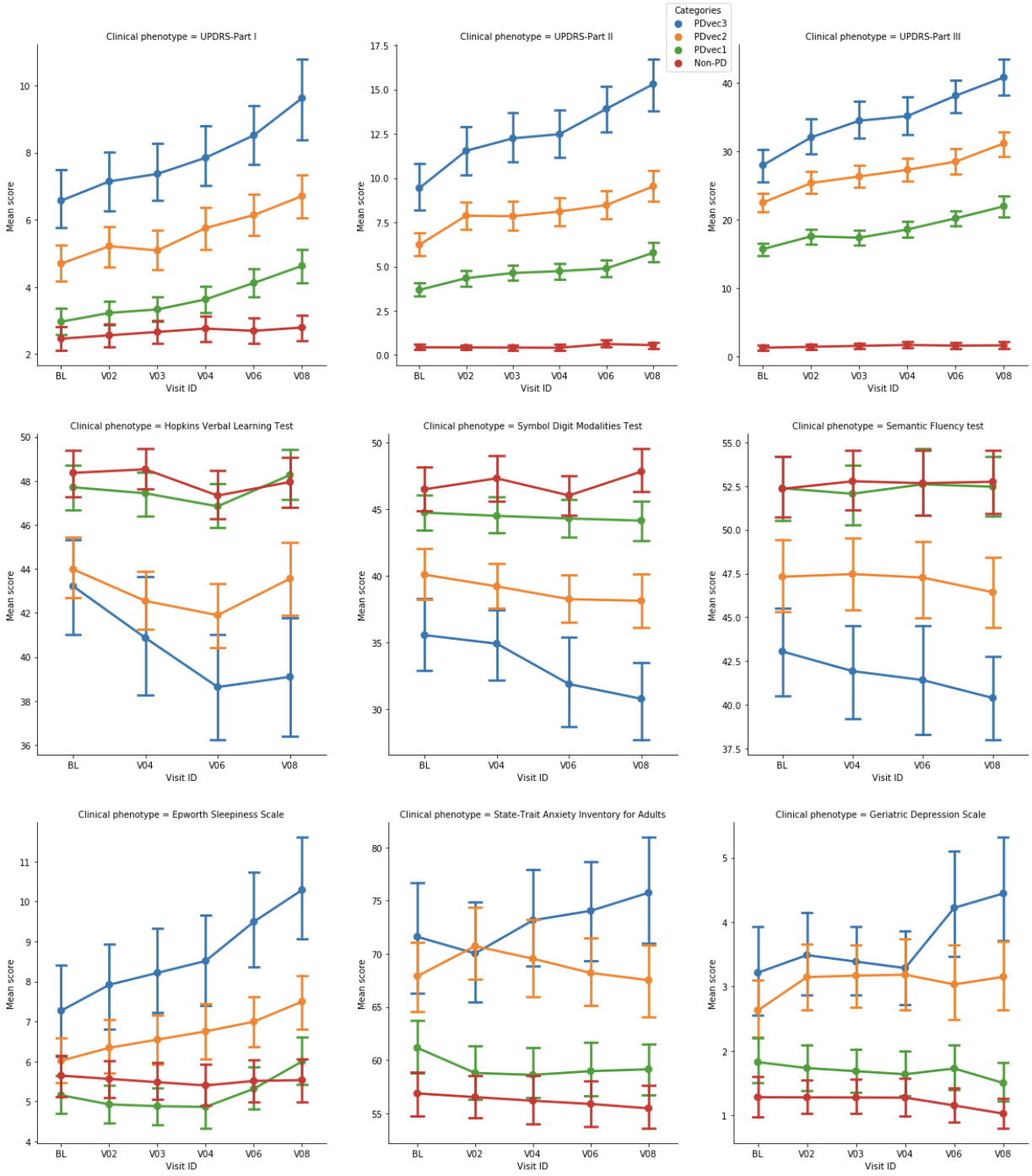

Figure 2.6: Shows the progression of each PD subtype over time. The top three graphs show the increased values of motor, sleep, and cognitive dimensions reflect the increase in disturbance overtime. The rest of the graphs demonstrate the actual clinical values of each subtype overtime for UPDRS-Part I, Part II, Part III, as well as Hopkins Verbal Learning Test, Symbol Digit Modalities Test, Semantic Fluency test, Epworth Sleepiness Scale, State-Trait Anxiety Inventory for Adults, and Geriatric Depression Scale. BL: Baseline. V03: visit number 3 after 18 months. V04: visit number 4 after 24 months. V06: visit number 6 after 36 months. V08: visit number 8 after 48 months.

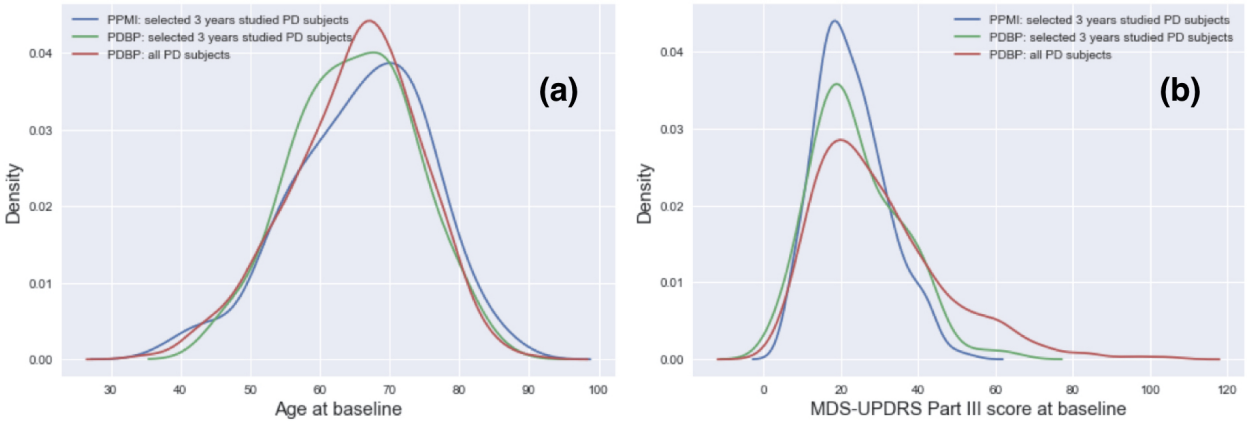

Figure 2.7: Kernel Density Estimation (KDE) analysis of Age and MDS-UPDRS Part III (objective motor symptom examination by a trained neurologist) in PPMI and PDBP cohorts. (a) shows the density of Parkinson's participant's age in the 3-years PPMI, PDBP, and 3-years PDBP datasets, and (b) shows the distribution of Parkinson's participant's MDS-UPDRS Part III at baseline in the 3-years PPMI, PDBP, and 3-years PDBP datasets. The three density functions in both figures are similar showing the validity of statistical replication.

Table 2.1: two-sample t-test for quantified replication cohort validation analysis. PPMI vs. PDBP (selected participants with 3 years data).

Parkinson participant's distribution of cohorts	***t-value***	***p-value***
Age	0.9244736	0.3557441
UPDRS Part III	-1.162273103	0.245751729

Following the data-driven organization of subjects into progression subtypes and clustering them into three subtypes, we developed three models to predict patient progression class after 48 months based on varying input factors: (a) from baseline factors, and (b) from baseline and year 1 factors. Figure 2.9a and Figure 2.9b show the ROC (Receiver Operating Characteristic) curves of our multi-class supervised learning predictors. We correctly distinguish patients with Parkinson's disease based on baseline only input factors and predict their 48-month prognosis with an average AUC of 0.93 (95% CI: 0.96 $\pm$ 0.01 for PDvec1, 0.87 $\pm$ 0.03 for PDvec2, and 0.96 $\pm$ 0.02 for PDvec3). The predictor built on baseline and year 1 data performs even better with an average AUC of 0.956 (95% CI: 0.99 $\pm$ 0.01 for PDvec1, 0.91 $\pm$ 0.03 for PDvec2, and 0.97 $\pm$ 0.02 for PDvec3). The increased accuracy is due to the availability of more information about a subject. This approach is also practical in a clinical setting, as physicians will provide a better prognosis of patients after a year follow up.

The predictive model was also analyzed and enhanced by using a feature extraction

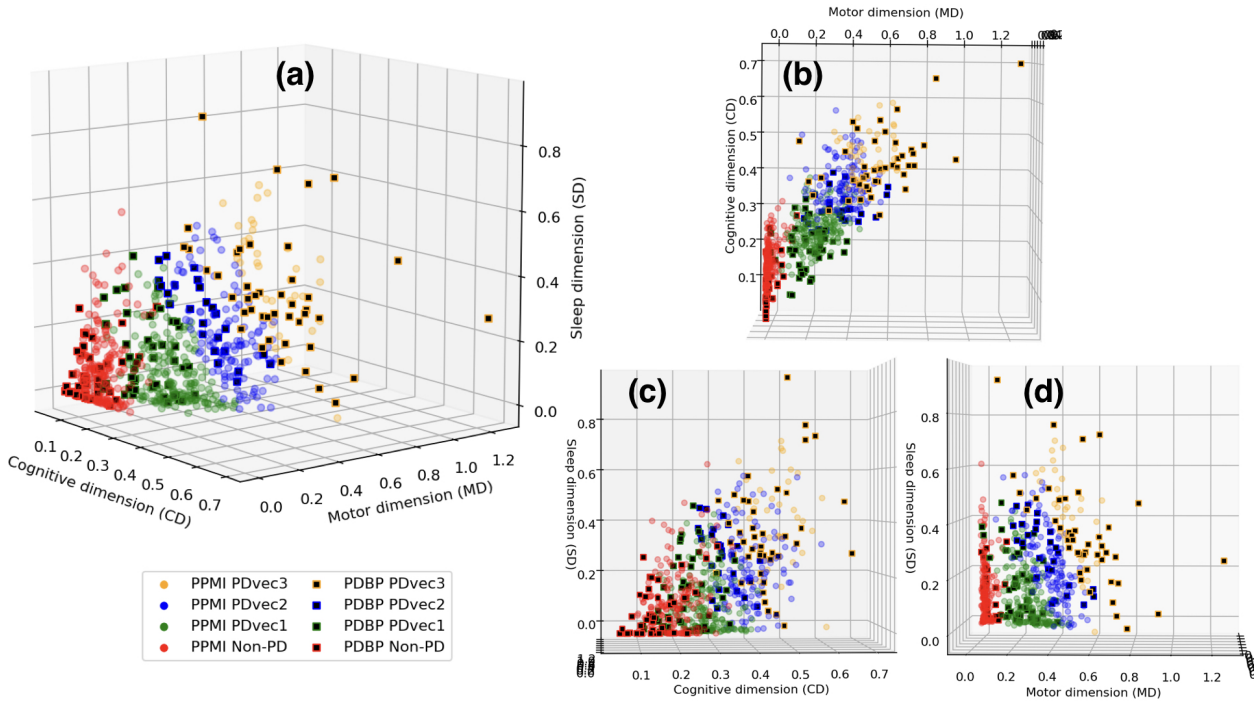

Figure 2.8: Shows the identified subtypes in the independent PDBP cohort using the model developed on the PPMI dataset. Similar PDBP and PPMI subtypes in terms of progression. (a) shows the view of all three dimensions, (b) view of the motor and cognitive dimensions, (c) view of cognitive and sleep dimensions, and (d) view of sleep and motor dimensions.

method: Recursive Feature Elimination. For the predictive model based only on baseline factors, out of 140 clinical parameters, 52 were identified to be the significant contributors (Table 2.2 for list and detail). Essentially, having only 52 parameters will provide us with the highest prediction accuracy. For the predictive model on baseline and year 1 factors, incorporating 66 parameters out of 250 (not all factors were measured at both baseline and first-year) provided us with the highest prediction accuracy (Table 2.2 for list and detail). From these 66 parameters, 34 are baseline measurements and the same as baseline predictor, three new baseline measurements, and 29 measurements from the first year.

Besides the cross-validation of predictive models in the PPMI cohort, we have also validated the accuracy of the predictive model in the independent PDBP cohort. The predictive model trained on the PPMI baseline data correctly distinguished PDBP patients with AUC of 0.787 (ROC curves in Figure 2.9c). The replicated predictive model performs very well for PDvec1 and PDvec3 (AUC of 0.90 and 0.89, respectively). However, due to the small sample size, the predictive model does not predict as well on PDvec2 (AUC of 0.57). There are fewer samples that make up the PDvec2 cluster in the replication cohort, and it has been

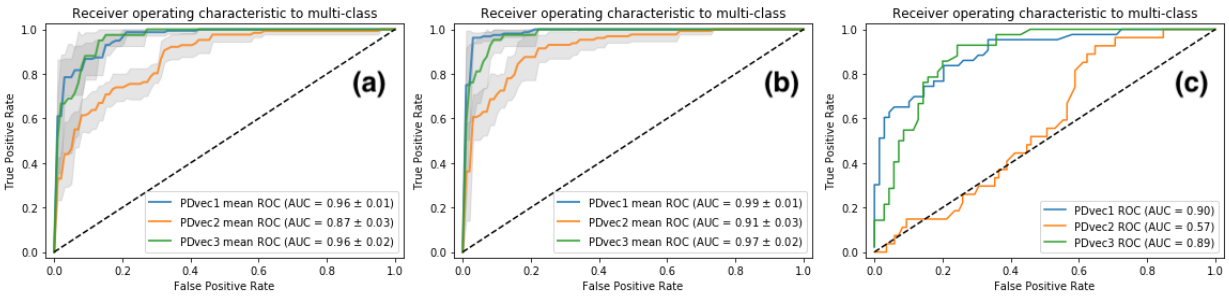

Figure 2.9: Shows the performance of Parkinson's disease progression prediction models. (a) The ROC for the predictive model at baseline developed on the PPMI cohort evaluated using five-fold cross-validation. (b) The ROC for the predictive model developed on the baseline and first-year data of the PPMI cohort evaluated using five-fold cross-validation. (c) The ROC for the predictive model developed on the PPMI baseline and tested on the PDBP cohort.

more accessible for the predictive model to predict the more extreme subtypes (i.e., PDvec1 and PDvec3). Despite the smaller sample size of the PDBP cohort, the results strongly validate our previous observations of distinct, computationally discernible subtypes within the PD population. This finding indicates that our methodology is robust, and our unique progression analysis and clustering approach is resulting in the same clusters.

Table 2.2: Summary of clinical parameters with significant contributions to the prediction models. Table lists significantly contributing clinical parameters based on baseline examination tests (BL) or based on the baseline with year-1 (BL + Y1) test items. Abbreviations: EPS, Epworth Sleepiness Scale; HVLT, Hopkins Verbal Learning Test; LNS, Letter-Number Sequencing; MDS-UPDRS, Movement Disorder Society Revision of the Unified Parkinson's Disease Rating Scale; MoCA, Montreal Cognitive Assessment; RBDSQ, REM Sleep Behavior Disorder Screening Questionnaire; QUIP, Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease; SCOPA-AUT, Assessment of Autonomic Dysfunction; STAI, State-Trait Anxiety Inventory.

Begin of Table 2.2

Clinical Scale Specific Test Item(s) (Parameter Name)	**Model**	
	Baseline	**Baseline + Year 1**
MDS-UPDRS Part I		
1.3 Depressed mood (NP1DPRS)	+	+ (BL)
1.4 Anxious mood (NP1ANXS)	+	+ (BL)
1.7 Sleep problems (NP1SLP)	+	+ (Y1)
1.9 Pain and other sensations (NP1PAIN)	+	+ (Y1)

Table 2.2 (cont.)

Clinical Scale Specific Test Item(s) (Parameter Name)	**Model**	
	Baseline	**Baseline + Year 1**
1.11 Constipation problems (NP1CNST)	+	+ (Y1)
1.13 Fatigue (NP1FATG)	-	+ (Y1)
MDS-UPDRS Part II		
2.2 Saliva and drooling (NP2SALV)	-	+ (Y1)
2.4 Eating tasks (NP2EAT)	+	+ (Y1)
2.5 Dressing (NP2DRES)	+	+ (BL, Y1)
2.6 Hygiene (NP2HYGN)	+	-
2.7 Handwriting (NP2HWRT)	+	+ (BL, Y1)
2.8 Doing hobbies and other activities (NP2HOBB)	-	+ (Y1)
2.11 Getting out of bed, car, deep chair (NP2RISE)	+	+ (Y1)
MDS-UPDRS Part III		
3.1 Speech (NP3SPCH)	+	+ (BL, Y1)
3.2 Facial expression (NP3FACXP)	+	+ (BL, Y1)
3.3a Rigidity neck (NP3RIGN)	+	+ (Y1)
3.3b Rigidity RUE (NP3RIGRU)	+	+ (BL, Y1)
3.3d Rigidity RLE (NP3RIGRL)	-	+ (Y1)
3.3e Rigidity LLE (NP3RIGLL)	+	-
3.4a Finger tapping right hand (NP3FTAPR)	+	+ (BL)
3.5b Hand movements left hand (NP3HMOVL)	+	+ (Y1)
3.6a Pronation-supination right hand (NP3PRSPR)	+	+ (BL)
3.6b Pronation-supination left hand (NP3PRSPL)	+	+ (Y1)
3.7b Toe tapping left foot (NP3TTAPL)	-	+ (Y1)
3.8a Leg agility right leg (NP3LGAGR)	+	+ (BL)
3.8b Leg agility left leg (NP3LGAGL)	+	+ (BL, Y1)
3.9 Arising from chair (NP3RISNG)	+	+ (BL)
3.13 Posture (NP3POSTR)	+	-
3.14 Global Spontaneity of movement (NP3BRADY)	+	+ (Y1)
3.16b Kinetic tremor left hand (NP3KTRML)	+	+ (BL, Y1)
3.17c Rest tremor amplitude RLE (NP3RTARL)	-	+ (Y1)
3.17d Rest tremor amplitude LLE (NP3RTALL)	-	+ (Y1)
3.21 Hoehn and Yahr stage (NHY)	+	+ (BL)
MoCA		
Naming total score (Naming)	+	+ (BL)

Table 2.2 (cont.)

Clinical Scale Specific Test Item(s) (Parameter Name)	**Model**	
	Baseline	**Baseline + Year 1**
Language total score (Language)	-	+ (BL)
Delayed recall total score (Delayed_recall)	+	+ (BL)
Abstraction (MCAABSTR)	+	+ (BL)
MoCA total score (MCATOT)	+	+ (BL)
HVLT		
Immediate Recall Trial 1 (HVLTRT1)	+	+ (BL)
Immediate Recall Trial 3 (HVLTRT3)	+	+ (BL)
Delayed Recall (HVLTRDLY)	+	+ (BL)
Recognition (HVLTREC)	+	+ (BL)
Recognition false positives, related (HVLTFPRL)	+	+ (BL)
LNS		
LNS-Sum questions 1-7 (LNS_TOTRAW)	+	+ (BL)
QUIP		
Think having issue with sex behavior (TMSEX)	+	+ (BL)
Think having issue with eating behavior (TMEAT)	+	+ (BL)
RBDSQ		
Dreams frequently have aggressive or action-packed content (DRMAGRAC)	+	+ (BL)
Know my arms and legs move when asleep (SLPLMBMV)	+	+ (Y1)
I (almost) hurt my bed partner or myself (DLPINJUR)	+	-
In my dreams: speaking, shouting, swearing (DRMVERBL)	+	+ (BL, Y1)
In my dreams: gestures, complex movements useless during sleep (DRMUMV)	+	-
In my dreams: things fell down around the bed (DRMOBJFL)	+	+ (Y1)
It happens that my movements awake me (MVAWAKEN)	-	+ (BL)
My sleep is frequently disturbed (SLPDSTRB)	+	+ (BL)
Disease of nervous system: stroke (STROKE)	+	+ (BL)
Disease of nervous system: depression (DEPR)	+	+ (BL, Y1)
EPS		
Doze off or fall asleep while watching TV (ESS2)	-	+ (BL)
SCOPA-AUT		

Table 2.2 (cont.)		
Clinical Scale Specific Test Item(s) (Parameter Name)	**Model**	
	Baseline	**Baseline + Year 1**
Had difficulty swallowing or have choked + Has saliva dribbled out of your mouth + Has food become stuck in your throat (Gastrointestinal_upper)	+	+ (Y1)
Have feeling during meal that you were full very quickly + Had problems with constipation + Had to strain hard to press stools + Had involuntary loss of stools (Gastrointestinal_lower)	+	+ (BL)
Semantic Fluency		
Total number of animals (VLTANIM)	+	+ (BL)
Total number of vegetables (VLTVEG)	+	+ (BL)
STAI		
Anxiety state score (A_state)	+	+ (Y1)
End of Table 2.2		

In summary, we have mined data to identify three clinically-related constellations of symptoms naturally occurring within our longitudinal data that summarize PD progression (41.6%, 29.5%, 28.9% variance loadings) comprised of factors relating to motor, sleep and cognitive. We also utilized supervised learning methods to identify the most informative factors across these symptomatologies to predict the velocities of decline for each patient relative to matched healthy controls with excellent accuracy (>90% after cross-validation) from baseline clinical data.

2.4 DISCUSSION

Prediction of disease and disease course is a critical challenge in the care, treatment, and research of complex heterogeneous diseases. Within PD, meeting this challenge would allow appropriate planning for patients and symptom-specific care (for example, to mitigate the chance of falls, identifying patients at high risk for cognitive decline or rapid progression, etc.). Perhaps even more importantly, at this time, prediction tools would facilitate more efficient execution of clinical trials. With models predicting a patient-specific disease course, clinical trials could be shorter, smaller, and would be more likely to detect smaller effects; thus, decreasing the cost of phase 3 trials dramatically and essentially reducing the exposure of pharmaceutical companies in a typically expensive and failure-prone area.

We previously had considerable success in constructing, validating, and replicating a model that allows a data-driven diagnosis of PD and the differentiation of PD-mimic disorders, such as those patients who have parkinsonism without evidence of dopaminergic dysfunction [72]. We set out to expand this work by attempting to use a somewhat similar approach to 1) define natural subtypes of disease; 2) attempt to predict these subtypes at baseline; and 3) to identify progression rates within each subtype and project progression velocity.

While the work here represents a step forward in our efforts to sub-categorize and predict PD, much more needs to be done. The application of data-driven efforts to complex problems such as this clearly works; however, the primary limitation of such approaches is that they require large datasets to facilitate model construction, validation, and replication. To achieve the most powerful predictions, these datasets should include standardized phenotype collection and recording. Collecting such data is a challenge in PD, with relatively few cohorts available with deep, wide, well-curated data. Thus, a critical need is the expansion or replication of efforts such as PPMI or PDBP, importantly with a model that allows unfettered access to the associated data; the cost associated with this type of data collection is large, but these are an essential resource in our efforts in PD research.

2.5 SUMMARY

In this study, we addressed the complexities of Parkinson's disease. We integrated unlabeled, multi-modal, and longitudinal data. The longitudinal data had a long-term nature, and we were interested in capturing the overall pattern of the individual's trajectories. Vectorization and NMF methods were the most successful approach for extracting long-term trajectories. Using comprehensive multi-modal data helped us develop an embedded space. This space was crucial for understanding the trajectories and dimensions in which the individuals traverse. Having this easily interpretable space, we were able to use GMM unsupervised learning approach to identify new subtypes of the disorder based on disease progression. We also provided an in-depth analysis of these subtypes. Furthermore, we developed predictive models for early diagnosis, prognosis, and clinical trial stratification.

This work provides data-driven subtypes in distinct progression stages of PD and discusses an approach to predict the future rate of progression years from baseline using longitudinal clinical data. Predicting disease progression serves as a paramount challenge in the therapy and cure of several elaborate diseases. This study is a step forward towards designing

sophisticated machine-learning paradigms to facilitate early diagnosis of PD progression. Predicting PD progression rates would lead to better patient-specific attention by recognizing at an early stage the patients with a swift rate of progression. The proposed disease progression and trajectory prediction algorithms can help doctors and practitioners develop a methodical and organized course for clinical tests, which can be much more concise and effective in detection. These adaptations and modifications in clinics may help to diminish treatment and therapy costs for PD. Further, the capability to anticipate the trajectory of impending PD progression at the early stages of the disease is an advancement towards uncovering novel treatments for PD modification. The proposed analysis provides insights to inhibit or decelerate the progression of PD-related symptoms and subsequent deterioration in the characteristics of life that are accompanied by the disease.

CHAPTER 3: LEARNING DEMENTIA AND ALZHEIMER'S DISEASE

In this chapter, we review the work on *Predicting Alzheimer's disease progression trajectory and clinical subtypes using machine learning*[1]. Similar to the work on Parkinson's disease, we use an unsupervised learning approach to identify new subtypes of the disorder based on disease progression. We also provide an in-depth analysis of these subtypes. Furthermore, we develop predictive models for early diagnosis, prognosis, and stratification. This work previously appeared as [48, 49, 50, 51].

Alzheimer's disease (AD) is a common, age-related, neurodegenerative disease that impairs a person's ability to perform day to day activities. Diagnosing AD is difficult, especially in the early stages, many individuals go undiagnosed partly due to the complex heterogeneity in disease progression. This highlights a need for early prediction of the disease course to assist its treatment and tailor therapy options to the disease progression rate. Recent developments in machine learning techniques provide the potential to not only predict disease progression and trajectory of AD but also to classify the disease into different etiological subtypes.

The suggested work clusters participants in distinct and multifaceted progression subgroups of AD and discusses an approach to predict the progression stage from baseline diagnosis. We observe that the myriad of clinically reported symptoms summarized in the proposed AD progression space corresponds directly to memory and cognitive measures, classically been used to monitor disease onset and progression. The proposed work concludes notably accurate prediction of disease progression after four years from the first 12 months of post-diagnosis clinical data (Area Under the Curve of 0.92 (95% confidence interval (CI), 0.90-0.94), 0.96 (95% CI, 0.92-1.0), 0.90 (95% CI, 0.86-0.94) and 0.83 (95% CI, 0.77-0.89) for controls, high, moderate and low progression rate patients respectively). Further, we explore the long short-term memory (LSTM) neural networks to predict the trajectory of a patient's progression.

The machine learning techniques presented in this study may assist providers with identifying different progression rates and trajectories in the early stages of disease progression, hence allowing for more efficient and unique care deliveries. With additional information about the progression rate of AD at hand, providers may further individualize the treatment

[1]This research was assisted by Vipul Satone and Rachneet Kaur as documented in their thesis.

plans. The predictive tests discussed in this study not only allow for early AD diagnosis but also facilitate the characterization of distinct AD subtypes relating to trajectories of disease progression. These findings are a crucial step forward to early disease detection. Additionally, models can be used to design improved clinical trials for AD research.

3.1 INTRODUCTION

Alzheimer's disease (AD) is a progressive and age-associated chronic neurodegenerative disease affecting a patient's memory, intellectual skills, and other mental functions. It is the most common form of dementia. Research has shown that AD is a clinically heterogeneous condition, showing marked variations in terms of the symptoms constellations and disease progression rates. The clinical signs and symptoms of AD show marked variability in terms of patients' age, disease span, progression velocity, and types of memory, cognition, and depression-related features. After the age of 65, the prevalence of dementia doubles every five years and is known to increase exponentially after the age of 90 [92]. As dementia affects older people, with a growing life expectancy, it is becoming a crucial medical problem [93].

With no preventive interventions known, AD progression is a major concern for health care providers around the globe. Researchers have shown that AD pathological changes occur 20 years or earlier before the actual disease symptoms manifest [94, 95, 96, 97, 98, 99]. In the absence of any cure or disease-modifying treatment for this disabling disease, current treatment strategies are limited to supportive, symptomatic care [100, 101]. Delay in the diagnosis of AD is often due to the disease complexity, with no clear identifying early diagnostic criteria available for providers [102]. A major challenge for AD prediction is the presence of inherent phenotypic diversity in the AD population. Hence, the idea of personalized clinical care with individualized risk, progression, and prediction related patient advice in AD is narrow. Additionally, there are ramifications in clinical trial design when considering the high heterogeneity of disease manifestation and progression. Predicting disease progression trajectories at an early stage is crucial for the design of clinical trials and the development of disease-modifying treatment strategies.

For the treatments to be most effective, the AD therapy regimen must likely begin before the notable downstream damage [103]. Simply put, early AD detection is a likely scenario to make the greatest therapeutic gains. Patients diagnosed with mild cognitive impairment (MCI) at study baseline are at a higher risk for progression to dementia, but not all patients end up developing AD [104]. Research has been done to detect AD in patients with MCI

or predict the early stage of AD using cerebrospinal fluid (CSF) [105, 106], while others [107] have used psychometric and imaging data for predicting the progression of dementia in patients with amnestic MCI. In an implementation of a multi-class classifier using clinical and magnetic resonance (MR) brain images to classify controls, MCI, and AD patients, an accuracy of 79.8% was achieved [108]. Less research has been done on using clinical data and predicting the AD progression rate. Recently, we have used machine learning to classify Parkinson's disease (PD) patients into three different sub-categories with highly predictable progression rates [43]. We explored variations in onset and progression velocity and observed clusters of the motor, cognitive, and sleep disturbance related features using the clinical data. In this work, we extend our latest approach by applying it to the clinical features of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [109].

3.1.1 Goals and Contributions

This work was designed to cluster AD patients into distinct progression groups and to predict the progression trajectory at an early baseline period. Dimensionality reduction via non-negative matrix factorization (NMF) was used to define an *ADNI progression space* for the AD summarizing myriad clinical measures across multiple time points. By applying unsupervised machine learning, namely, the Gaussian mixture model (GMM) on the extensive clinical observations available in the ADNI dataset, we algorithmically parsed the progression space for the AD into three clinical subtypes, defined as *low*, *moderate (medium)* and *high* disease progressors. Our analysis found that clinically related symptoms corresponding to memory and cognition make up the AD progression space. Clinical data collected at baseline (study entry), after 6 and 12 months, is used to predict memory and sleep decline after 24 and 48 months from baseline. We validated our models through five-fold cross-validation to obtain a robust prediction of memberships into these progression subtypes. Along with traditional machine learning methods, the long short-term memory (LSTM) neural networks were also used to predict disease progression rates (control, low, moderate, and high) after 24 and 48 months from baseline. The described methodologies may lead a step forward towards the development of personalized clinical care and counseling for patients, hopefully reducing AD therapy costs in the future. Also, we attempt to describe the trajectory of AD progression via LSTM networks.

Further, we examine the reversion instances of AD captured in the constructed progression space, the correlation of Apolipoprotein Eε4 (APOEε4) compound genotype with cognitive performance and interactions between certain selective features associated with AD and the

constructed progression space later in the discussion section of the paper. These observations provide a promising understanding of AD characteristics useful for devising novel disease modification therapies. The proposed analysis provides a potential understanding towards restraining AD-related symptoms and consequent deterioration in the life of the patients. We believe that the advancement of the discussed prediction models has the potential to impact clinical decision making and improve healthcare resource allocation in AD significantly.

3.2 METHODS

The data analysis pipeline for this work was performed in Python 3.6 with the support of several open-source libraries (TensorFlow, scikit-learn, pandas, seaborn, etc.). To facilitate replication and expansion of this study, the Python code (including the entire data preprocessing and machine learning analysis) was made publicly available under GPLv3 as part of the supplementary information at `https://github.com/vipul105/Alzheimers_Disease_Progression`.

3.2.1 Study design and participants

Data used in the preparation of this article were obtained from the ADNI database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. The ADNI dataset involves participants from over 50 sites across North America and Canada. All participants and their study partners provided their consent, accepting their engagement for the data collection, and the study protocols for ADNI were approved by the Institutional Review Board. The ADNI study was carried out in phases, namely, ADNI 1 beginning in 2004, followed by ADNI GO in 2009 and ADNI 2 in 2011. These editions had different participants and data collection procedures, accounting for advancement in technologies. For more up-to-date information, see `www.adni-info.org`.

The eligibility criteria for the ADNI participants and further details on the protocol can be found at [109]. All participants went through comprehensive functional, cognitive, and clinical assessments and provided a blood sample for APOE genotyping at their baseline visit (study entry). These assessments and their status (control, MCI, and AD) were then

updated longitudinally at 6, 12, 18, 24, 36, and 48 months. In our analysis, predictions were made for each participant's AD stage after 24 and 48 months using up to 12 months of clinical data. The study consisted of 247 observations (with 123 (49.79%) females, the average age for all participants was 71.55 ± 6.79 years and 94.73% of them are of European ancestry) for prediction at the 48^{th} month and 453 observations (with 208 (45.92%) females, the average age for all participants is 72.32 ± 7.13 years, and 93.59% of them are of European ancestry) for prediction at the 24^{th} month. For observations corresponding to the 24^{th} month, mean age is 72.84 ± 6.09, 71.61 ± 7.47 and 72.92 ± 8.11 for controls, MCI and dementia patients respectively and for observations corresponding to the 48^{th} month, the mean age is 72.17 ± 6.67, 71.36 ± 6.67 and 70.34 ± 7.42 for controls, MCI and dementia patients respectively.

The total scores and subscores from the following commonly collected cognitive, functional, and longitudinal clinical data elements were used in the proposed work:

i Montreal Cognitive Assessment [74]

ii Clinical dementia rating [110]

iii Neuropsychiatric inventory questionnaire [111]

iv Neuropsychological battery [112]

v Mini-mental state exam (MMSE) [113]

vi Geriatric Depression Scale [82]

vii Everyday cognition - study partner [114]

viii Everyday cognition - participant [115]

ix Functional assessment questionnaire (FAQ) [116]

We considered a total of 145 clinical variables (features) from the above-mentioned assessments for our analysis.

3.2.2 Procedures and statistical analysis

Only the observations which had data recorded for all the considered tests were taken into account. To construct the AD progression space, we used readings taken at baseline and on visits after 6 and 12 months from the baseline.

3.2.3 ADNI progression space and the prediction model

We leverage the temporal information present in the data to manage missing data recordings. Missing values were imputed using linear interpolation based on the past visit readings for the feature, therefore avoiding any influence of other observations during data imputations. After the imputation, around 7% of the data was reduced. One hot encoding was used for categorical variables whenever required. Scaling the continuous features to a comparable range is necessary to avoid the influence of certain features over others. Min-max normalization was used to retain the progressions since the ADNI dataset in consideration is multi-modal. Furthermore, min-max normalization didn't affect categorical features. Figure 3.1 shows our detailed workflow pipeline followed during the analysis. To reduce the dimensionality of the dataset, NMF [86, 87] (with a rank of 2) was used on 582 observations with available data for baseline, visits after 6 and 12 months. We used NMF to deconstruct data into two matrices, namely progression vectors and the progression indicators, which correspond to the latent vectors. Progression vectors were used to construct the 2-dimensional (2D) ADNI progression space. This 2D space was then used to predict a participant's disease progression stage after 24 and 48 months from baseline. Progression indicators map the features in the original dataset to the progression space, via which we identified memory and cognitive decline as the two dimensions of the modeled AD progression space. The relative position along the x- and the y-axis represent worsening sleep or memory disorder.

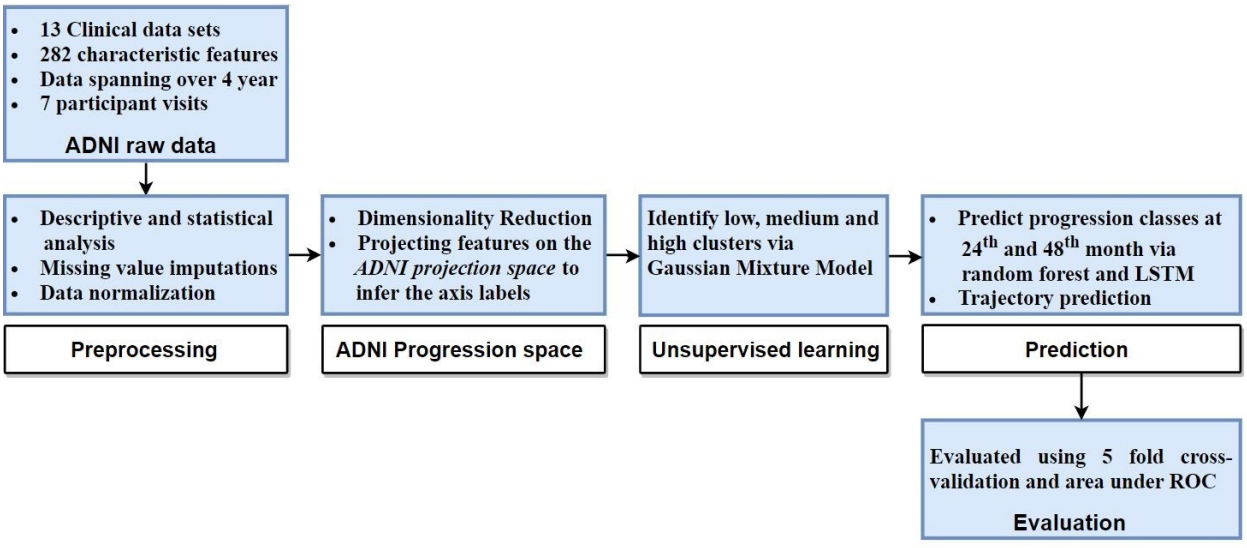

Figure 3.1: Workflow of analysis and model development.

Next, unsupervised clustering via GMM [88] was used to define the hidden subtypes within the MCI and dementia patients. GMM is an expectation-maximization algorithm

that maximizes the likelihood of observing the data, given the underlying parameters of the distribution. Bayesian information criterion (BIC) [90] was used to select the optimum number of underlying clusters for the GMM. BIC is a maximum likelihood estimate which tries to select the best model among the given set of candidates. In all the scenarios, three optimum number of clusters, defined as *low*, *moderate (medium)* and *high* progression rates, were attained. After obtaining the AD progression space and classifying MCI and dementia patients into different progression groups, the performance of various supervised learning classifiers (namely ensemble random forest, linear discriminant analysis, Naive Bayes, adaptive boosting, nearest neighbors, logistic regression and decision trees) were compared to predict a participant's progression stage after 24 and 48 months from baseline using readings up to 12 months. Two models were built a) Model 1: predict progression at 24^{th} month after baseline by using baseline and first-year factors b) Model 2: predict progression at 48^{th} month after baseline by using baseline and first-year factors. Recurrent neural network (RNN) architecture with LSTM was also used to predict the progression rates (control, low, moderate, and high) after 24 and 48 months from the baseline. The LSTM architecture had a single LSTM bidirectional layer connected to a fully connected layer. Cross entropy loss function was used at the output layer since it combines both logs of softmax and negative log-likelihood loss functions. Optimal parameters for the models were found to be a single hidden layer with 128 hidden units with a learning rate of 0.001 and a dropout probability of 0.2. Since our dataset size was limited in terms of the number of observations, five-fold cross-validation (CV) was used to evaluate the models. Among all the explored algorithms, a random forest classifier [91] gave the best five-fold CV accuracy. Hence, parameters for the random forest algorithm were fine-tuned using grid search (4800 iterations) and five-fold CV.

3.2.4 Model evaluation

Two different evaluation metrics were used for validating the clustering and prediction models. Sensitivity and specificity are measures of the proportion of positives that are correctly identified and negatives that are correctly identified, respectively. The plot of sensitivity on the y-axis and $(1 - \text{specificity})$ on the x-axis is called the area under the receiver operating characteristic (AUC of ROC) curve with a greater value representing a better clustering model. AUC of ROC was used to evaluate the clustering algorithms. Since this is a multi-class problem, one versus all approach was used to calculate the AUC for each class. Next, a five-fold CV was used to judge the performance of the proposed prediction models. The model was repeatedly trained on four parts, and accuracy for prediction was

calculated on the fifth part with a random selection of partitions each time.

3.2.5 Trajectory prediction

NMF was used to project the observations in progression space. LSTM was used to predict the position of patients in the progression space (trajectory prediction) at the 24^{th} and 48^{th} month using data collected at baseline and after 6 and 12 months from baseline. For this study, five separate projections were made using data until each visit and projections at the 24^{th} and 48^{th} month were predicted. A bidirectional LSTM with two layers consisting of 32 hidden units was trained for 50 epochs with a learning rate of 0.001 and a batch size of 10 for the same. Since the projection was made in the 2D axis, the mean Euclidean distance was used to assess the performance of this model. Only the features which were present for all the first three visits (baseline, m06, and m12 visits) were considered for this study.

In the subsequent analysis, we study the share of different frequencies of APOEε4 variants for each progression subtype since APOEε4 genotype is known to be closely related to AD risk [117]. Further, we discuss the reversion from AD to MCI and MCI to control stage captured in the proposed progression space and correlation of a participant's AD progression stage with their age, educational status, APOEε4 gene, and other selective critical features.

3.3 RESULTS

We have two progression indicator vectors in the reduced 2D ADNI progression space. The features observed in the real data were correlated to the two axes of the progression space using the magnitude of coefficients observed in the progression indicator vectors. A higher magnitude corresponding to the first progression indicator vector will correlate the feature to the first axis and similarly for the second axis.

3.3.1 ADNI progression space

Progression indicator i.e., coefficient matrix obtained from the NMF, was used to find out the hidden features that each of the two axes of the reduced space represents. Progression indicator vectors represent latent features of the reduced progression space. Progression indicator coefficients for each feature are plotted in Figure 3.2, and they are separated by drawing a line with slope 1. Features that occur below this separating line were associated with cognitive decline (x-axis) in the AD projection space, and features that lie above the

Table 3.1: Results for model 1 and model 2.

Progression Month	**Accuracy (95% CI)**	**AUC (95% CI)**			
		control	**low**	**moderate**	**high**
M24	84.98 ± 5.97	0.94 ± 0.03	0.90 ± 0.05	0.94 ± 0.03	0.98 ± 0.01
M48	79.75 ± 4.25	0.92 ± 0.02	0.83 ± 0.06	0.90 ± 0.04	0.96 ± 0.04

line were associated with memory decline (y-axis) in the AD progression space. Features close to the separating line were not associated with any axis.

This transformed data was used to project the participant's disease progression stage at the 24^{th} (Figure 3.3) and the 48^{th} month (Figure 3.4).

Further, three different zones, namely low, moderate, and high progression rates, were identified in the MCI and dementia patients at 24^{th} and 48^{th} month, as depicted in Figure 3.5.

The progression rate of participants after the 24^{th} and 48^{th} months from the baseline were predicted using the random forest classifier. It gave the best five-fold cross-validation accuracy and area under ROC curve results for all the cases. For the 24^{th} month using baseline and 12 months of observations, AUC of 0.94 (95% CI, 0.91-0.97), 0.98 (95% CI, 0.97-0.99), 0.90 (95% CI, 0.85-0.95) and 0.94 (95% CI, 0.91-0.97) for controls, high, low and moderate progression rates were observed respectively. Prediction of progression at 48^{th} month using baseline and 12 months of observations yields AUC of 0.92 (95% CI, 0.90-0.94), 0.96 (95% CI, 0.92-1.0), 0.90 (95% CI, 0.88-0.92) and 0.83 (95% CI, 0.77-0.89) for controls, high, low and moderate progression rates respectively. In our implementation, the accuracy for the prediction of progression at the 24^{th} and 48^{th} month is 83.60% and 77.33% respectively (Table 3.1). Figure 3.6 depicts the ROC curves for 4 separate classes (controls, low, moderate and high progression rates) for the 24^{th} and 48^{th} month.

LSTM model was also used for the prediction of AD subtypes with low, medium, and high progression rates as well as controls. The accuracy of prediction of projection rates using LSTM is 75.91% and 81.77% for the 48^{th} and 24^{th} month respectively. The performance of the neural network did not match other more traditional methods because of small sequence length, a smaller number of features, and a limited number of participants in the dataset.

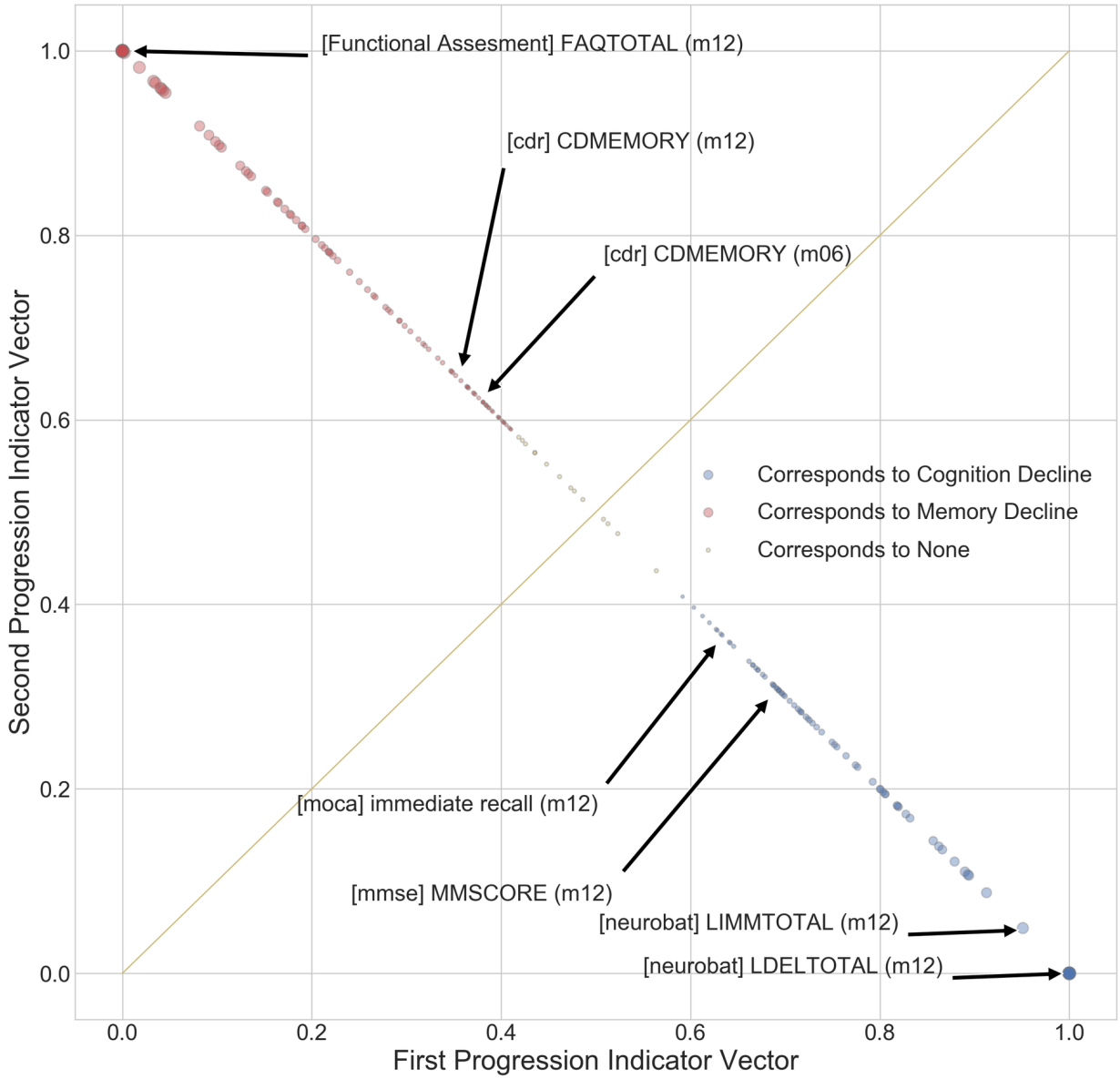

Figure 3.2: The plot of features in two dimensions using progression indicator vectors. Features in red correspond to memory decline and features in blue correspond to cognitive decline. Yellow line with a slope of 1, which separates the features into two categories, is drawn for reference.

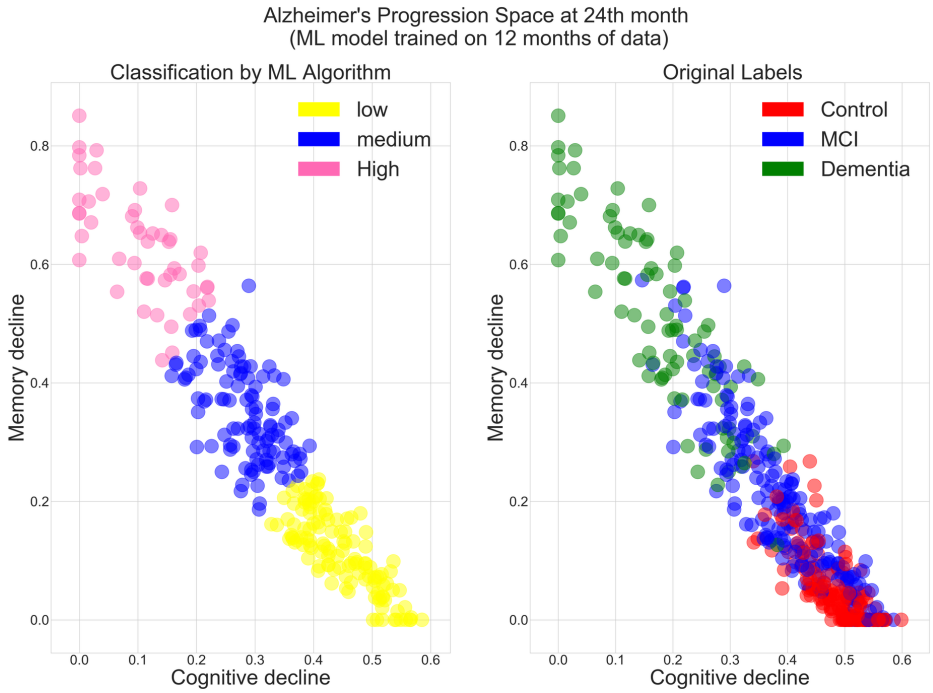

Figure 3.3: Comparison of 24$^{\text{th}}$ month machine learning based prediction and original labels. A total of 453 cases are projected in the AD progression space at the 24$^{\text{th}}$ month. Left: Machine learning based classification. Right: Colored with original labels.

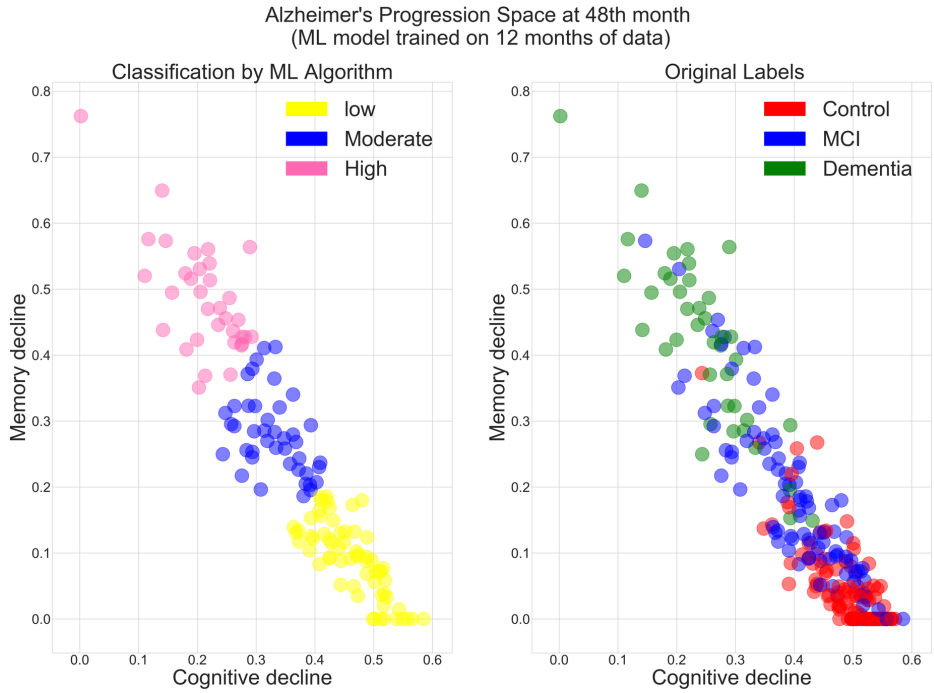

Figure 3.4: Comparison of 48$^{\text{th}}$ month machine learning based prediction and original labels. A total of 247 cases are projected in the AD progression space at the 48$^{\text{th}}$ month. Left: Machine learning based classification. Right: Colored with original labels.

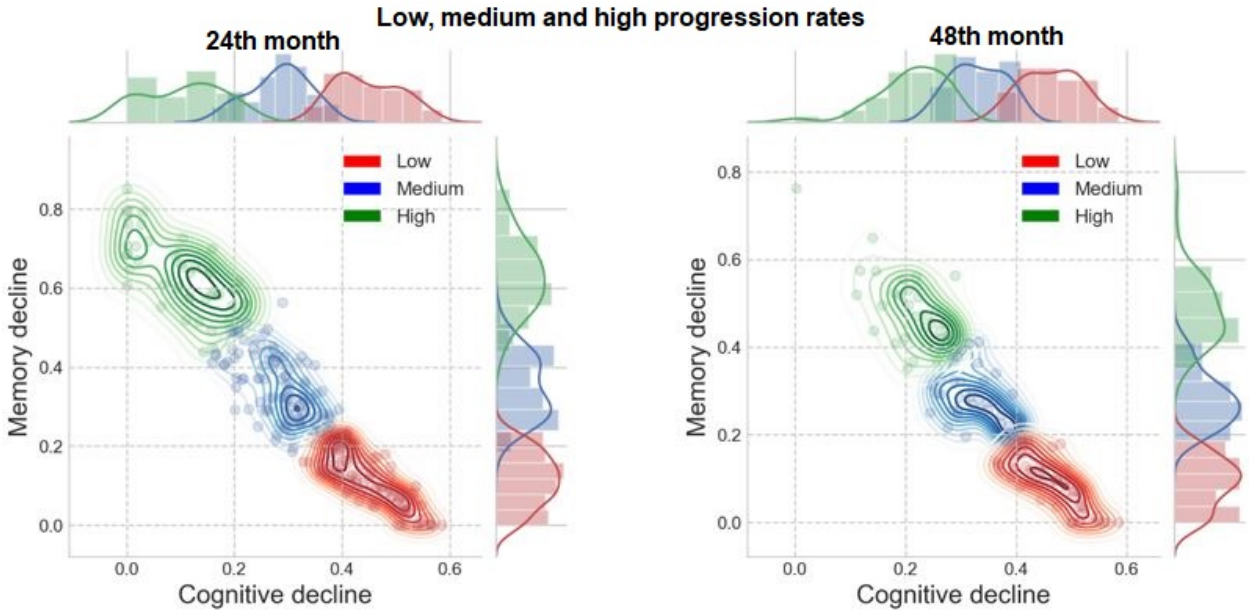

Figure 3.5: Three different progression rates are identified in MCI and dementia patients. Left: at the 24$^{\text{th}}$ month. The low, moderate and high progression rate zones are represented in red, blue and green respectively. Right: at the 48$^{\text{th}}$ month. The low, moderate and high progression rate zones are represented in red, blue and green respectively.

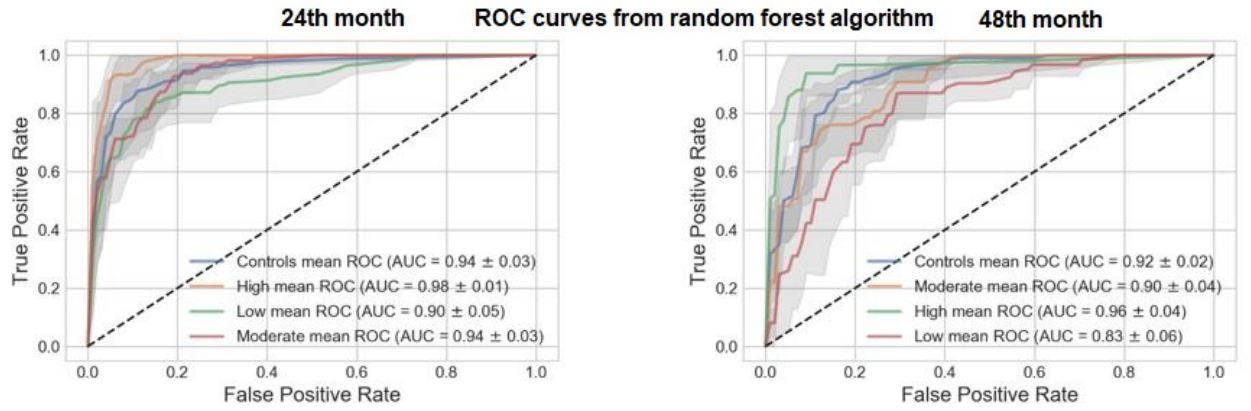

Figure 3.6: ROC of AD patient's progression rate after the 24$^{\text{th}}$ and 48$^{\text{th}}$ months based on the baseline data. Including the area under the ROC for 4 AD progression subtypes (controls, low, moderate and high progression rates). Left: The predictions for the disease stage at the 24$^{\text{th}}$ month were made using a random forest algorithm. Right: The predictions for the disease stage at the 48$^{\text{th}}$ month were made using random forest algorithm.

3.3.2 Trajectory prediction using LSTM networks

Since no therapies in AD modification are known yet, predicting the trajectory of AD progression at the early disease stage may offer a practical tool for refined clinical trials testing novel disease-modifying strategies. Figure 3.7 represents the predicted and projected values in the AD progression space for one of the participants.

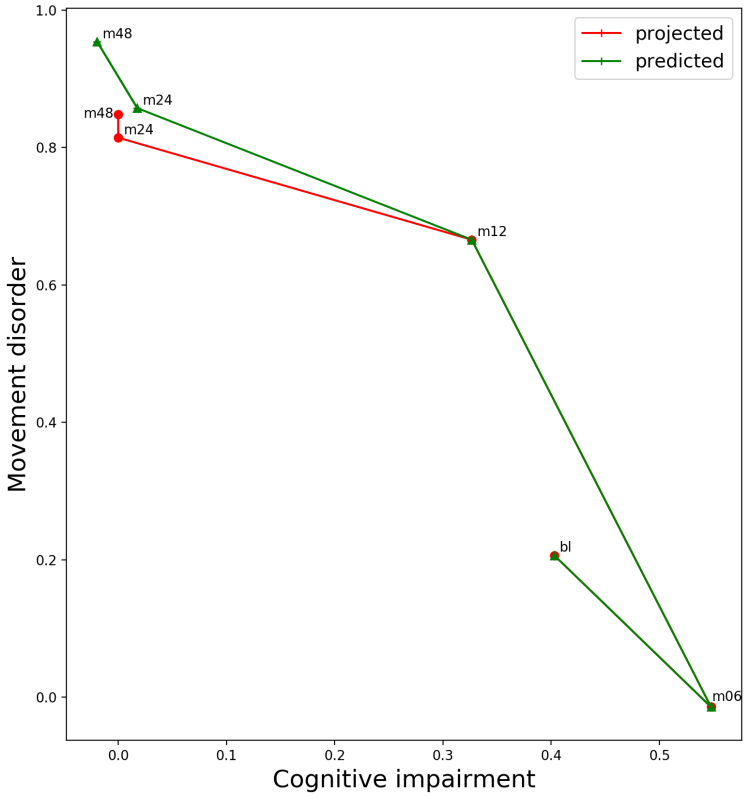

Figure 3.7: The predicted and projected trajectories of AD progression using LSTM for an AD patient.

NMF was used to project the data in progression space for each visit. The data from the first three visits were used to predict the position of an observation in progression space for the next two visits (visit at 48^{th} and the 24^{th} month). The mean Euclidean distance between projected and predicted test observations in the projection space (quantifying the model performance) for the 48^{th} and the 24^{th} month are 0.00206 and 0.00164 respectively. The number of participants in each cohort was limited, and not all the features were ascertained in all three (baseline, after 6 and 12 months) visits. These limitations affected the model performance.

Table 3.2: Number of MCI and dementia patients in each subtype.

	Low	**Moderate**	**High**
MCI	124	69	4
Dementia	1	42	43

3.3.3 Statistics of each progression subtype

The number of subjects with MCI, dementia, and controls in each subtype is shown in Table 3.2. Figure 3.8 shows their share percentage in each of the subtypes present after 24 months from baseline. As expected, the share of dementia patients is maximum at a high progression rate. The low progression rate has no dementia patient and contains only MCI patients. The moderate progression rate subtype is dominated by MCI patients, which covers around 83% of the observation in that subtype. A similar trend was observed after 48 months from baseline as well.

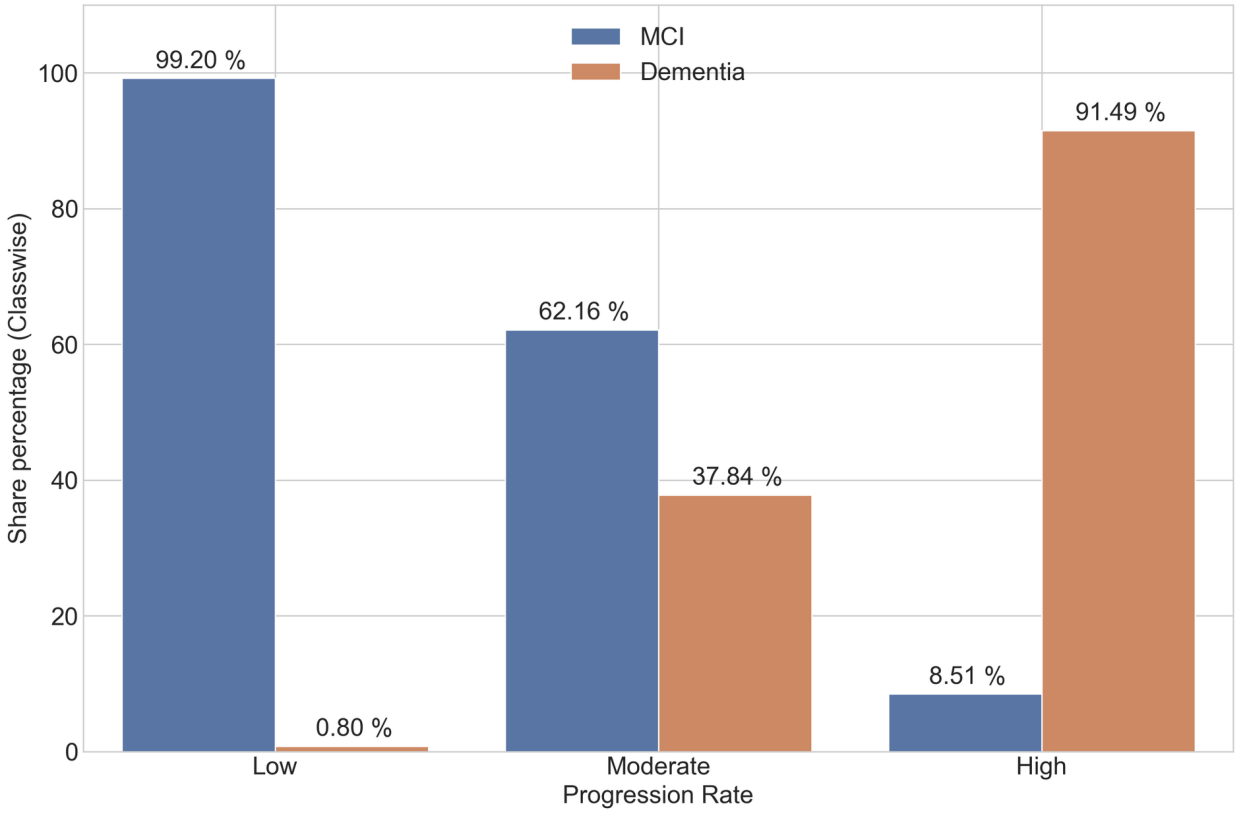

Figure 3.8: Percent share of controls, MCI and dementia patients for different subtypes present after 24 months from baseline. The share of MCI patients is decreasing with an increase in progression rate.

3.3.4 Percent share of APOEε4 variants for different subtypes

APOE has three common alleles, ε2, ε3, and ε4, of which the ε4 allele is closely associated with increased risk of AD [117]. The distribution of APOEε4 alleles for each progression rate subtype is shown in Figure 3.9 after 24 from the baseline. This figure illustrates that the progression rate increases with the number of APOEε4 alleles.

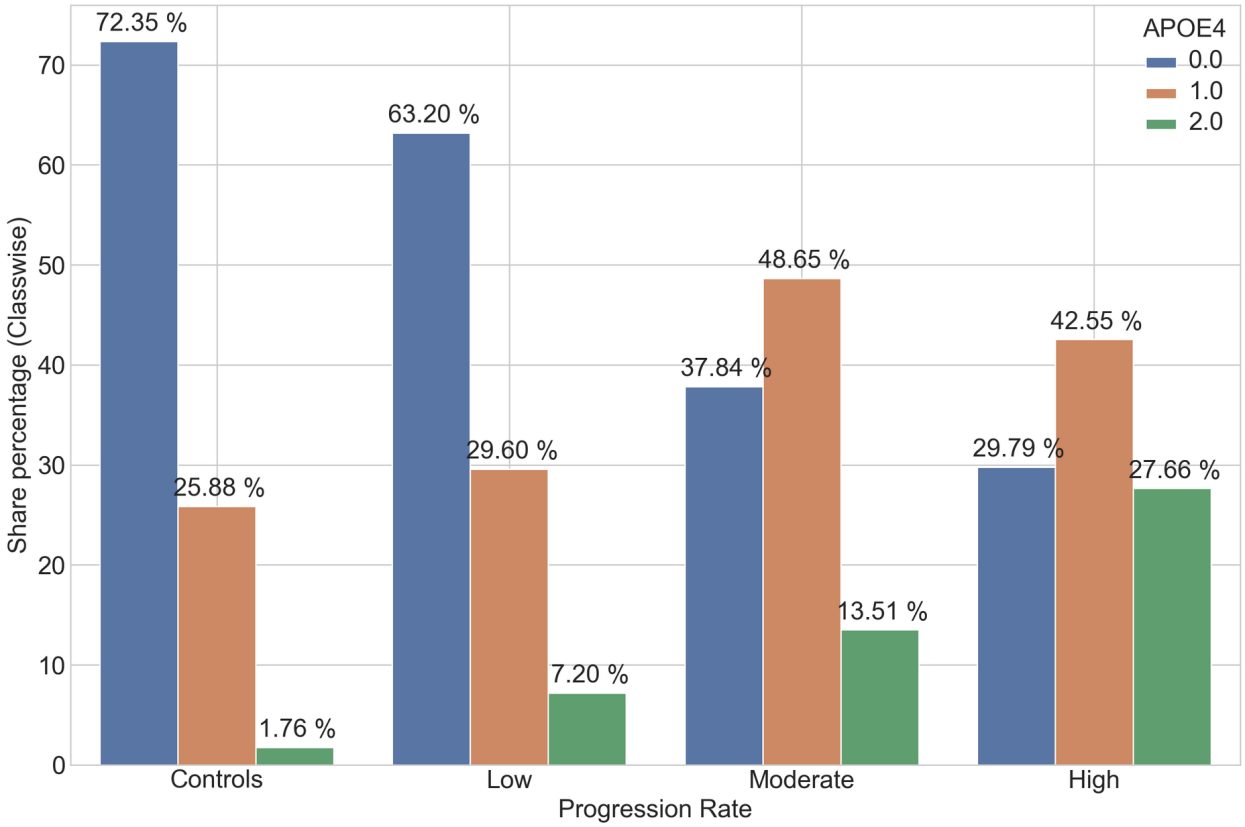

Figure 3.9: Percent share of APOEε4 alleles for different subtypes after 24 months from baseline. The share of 0 occurrences of APOEε4 alleles is decreasing with an increase in progression rate, whereas the share of 1 and 2 occurrences are increasing.

3.4 DISCUSSION

In this section, we study the reversion of AD captured in the constructed progression space, correlation between the APOEε4 genetic variants and participant's progression state, effects of aging on AD progression in controls, correlation of memory decline in AD patients with their educational and occupational attainments and distribution of projected dimensions (memory decline and cognitive decline) for each AD subtype, correlation between certain

selective features and AD progression rates in the following subsections. To conclude, we discuss some of the future directions to extend the proposed study.

3.4.1 AD progression space and disease reversion

Since ADNI is a longitudinal study, the disease state of patients is reassessed every 12 months. The clinical condition either deteriorated or stayed the same for most of the patients, but in rare instances, it reversed to a better state, i.e., some patients were observed moving from dementia to MCI or MCI to control stage. These observations were plotted to assess the robustness of the constructed progression space. Figure 3.10 plots these reversion cases at the 24^{th} and 48^{th} month. It can be verified from these figures that patients moving from dementia to MCI fall in the intermediate region between dementia and MCI (moderate progression rate region). Similarly, patients moving from MCI to control lie in the intermediate progression region between them. Thus, the progression space captures the reversion of the disease state.

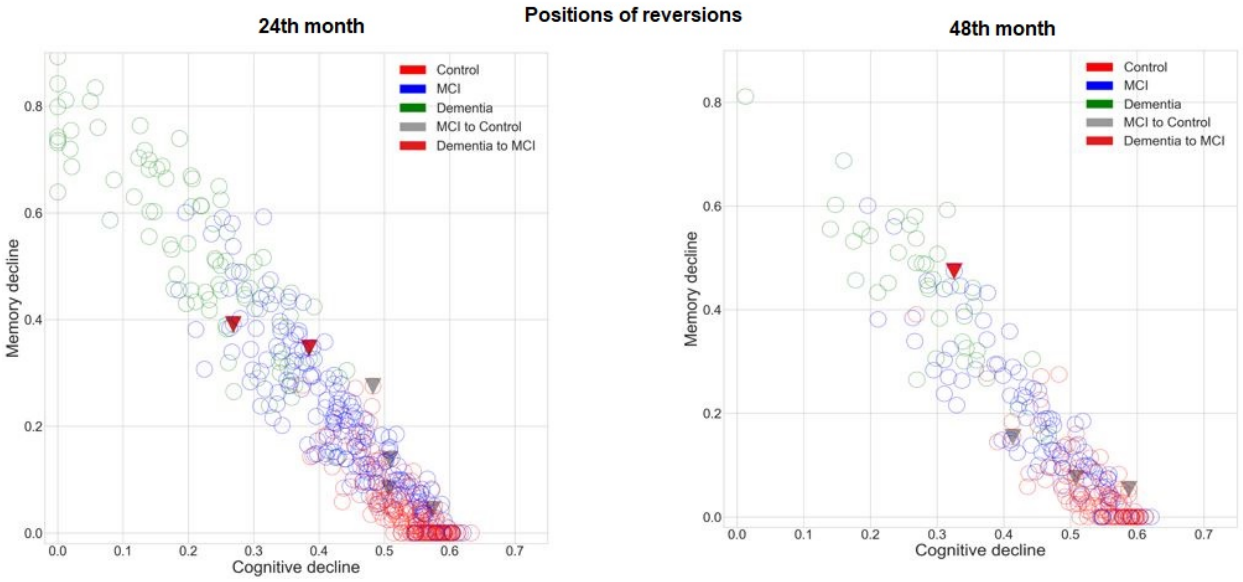

Figure 3.10: Label Reversions (only MCI to control and dementia to MCI). Left: The positions of the reversions at 24^{th} month relative to all other observations. Right: The positions of the reversions at 48^{th} month relative to all other observations.

3.4.2 AD progression states and APOEε4 allele counts

To understand the underlying biological patterns among patients in the progression space, we plotted the distribution of the APOEε4 alleles. Figure 3.11 projects the observations with

0 and 2 counts of APOEε4 variants on the AD progression space at 24^{th} and 48^{th} month. It is evident from these figures that observations with a 0 count for the APOEε4 allele are concentrated towards the low progression rate zone, whereas observations with two counts of APOEε4 allele are concentrated towards the moderate and high progression rate zones. This observation further validates the existing literature [118] and confirms a significant correlation between APOEε4 with cognitive performance.

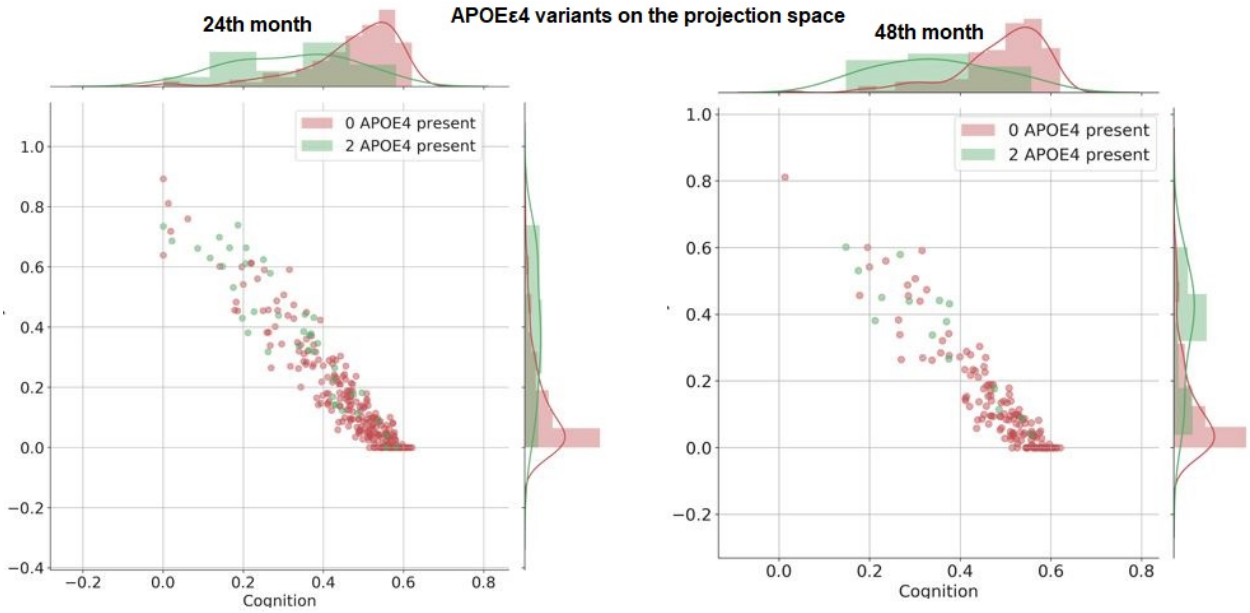

Figure 3.11: Projection of the number of APOEε4 variants on the projection space. x-axis and y-axis have visualizations of the distribution of APOEε4 alleles in those directions. Left: at 24^{th} month. Right: at 48^{th} month.

3.4.3 AD progression in controls and aging

In Figure 3.12, progression can also be seen in control observations at 24^{th} and 48^{th} month respectively, attributed to a decline in normal cognition and memory with increasing age of the participants. Since this decline is not severe, the observations do not lie in moderate or high progression rate zones. A simple clustering of observations into two clusters shows a stark difference in the mean age of the clusters. It is interesting to note that the mean age for the cluster, which is relatively close to the moderate progression rate zone is 74.59 years and the mean age of the cluster away from this zone is 72.15 years. Similarly, for the 48^{th} month, the mean age of the two clusters relatively close and away from the moderate progression zone are 71.53 and 73.51, respectively (Figure 3.12).

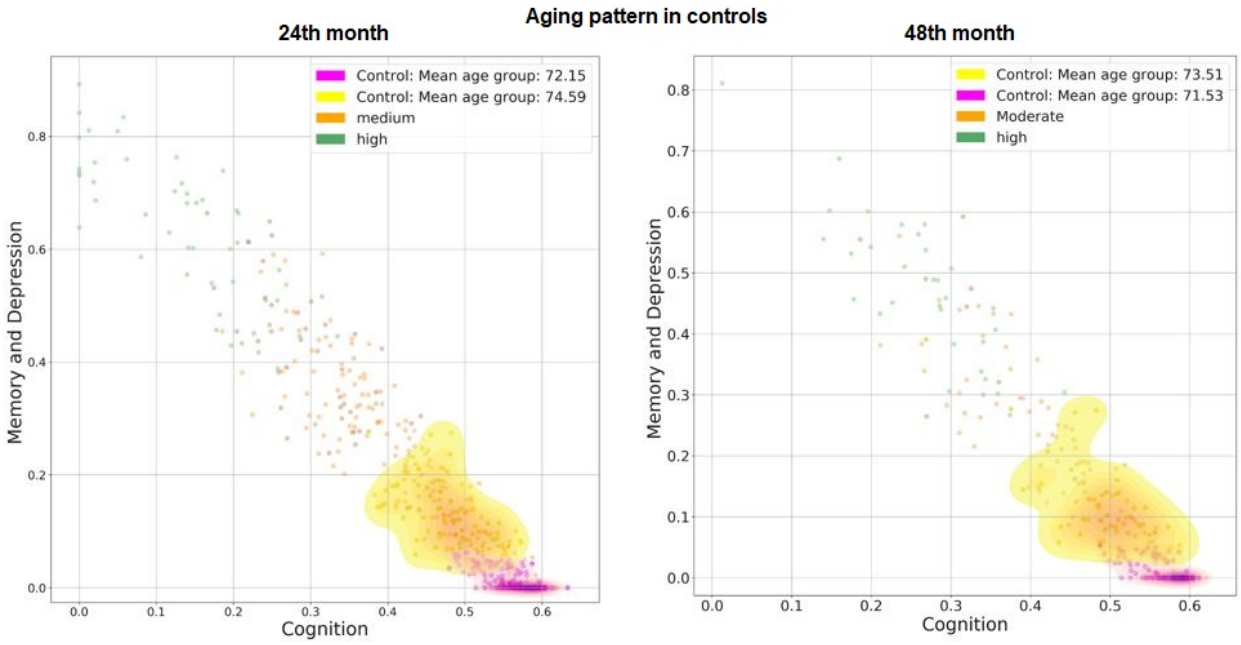

Figure 3.12: Aging pattern in controls. Mean age of cases in the two clusters of controls. Clusters represent an aging pattern in controls. The cluster near moderate progression rate zone has a high mean age than one which is away from it. Left: at 24^{th} month. Right: at 48^{th} month.

3.4.4 Memory decline in AD patients and educational acquirements

To further discover a generalized trend in the AD progression rate, a polynomial curve was fitted on the projected observations. BIC was used to find out the optimum degree of the fitting polynomial, which was observed to be three. As seen in Figure 3.13 (Left), the cubic curve fits the data in a linear fashion in the low progression region. However, it deviates slightly from this linear behavior in high and moderate progression region. The magnitude of the slope of the linear curve is 1.19 indicating a rapid memory decline as compared to cognitive. The slope of the progression for 200 most and least educated observations is shown in Figure 3.13 (Right). The magnitude of slope for the linear curve is 1.26 and 1.19 for highly educated and less educated patients, respectively. As the slope for highly educated patients is greater than the slope for less-educated patients, it can be inferred that there is a relatively rapid decline in memory of patients with higher education. A study on the links between education and memory decline in AD was carried out in [119]. The research concluded that memory declined more rapidly in AD patients with higher educational and occupational attainment. Thus, our results are further validated by these explorations done in the previous research [119].

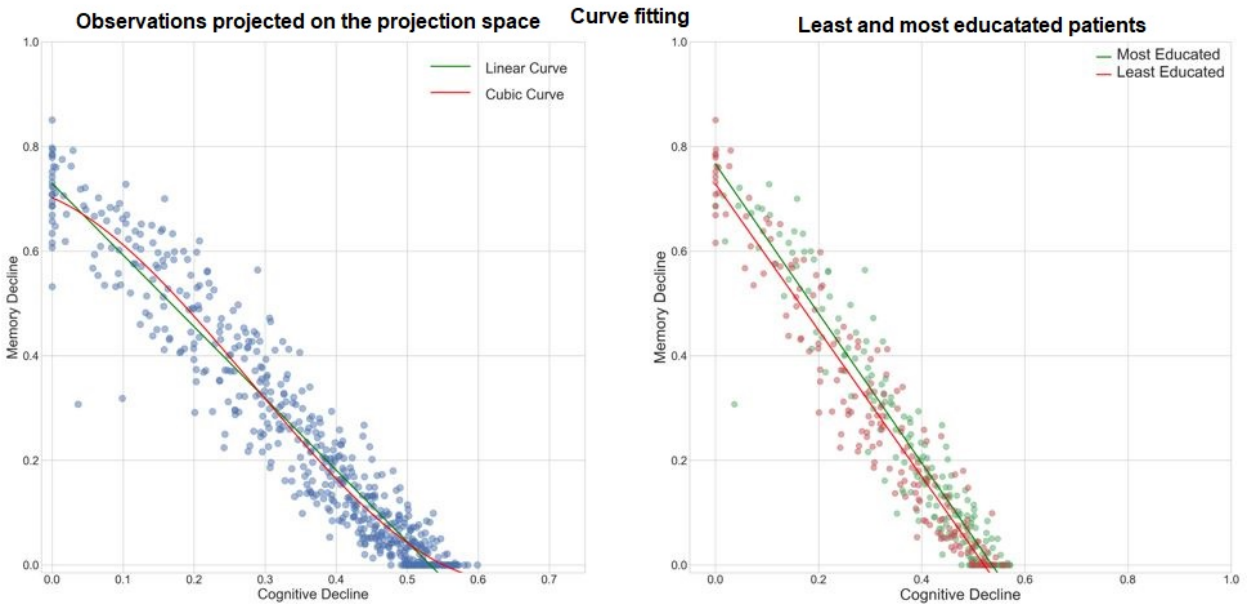

Figure 3.13: Memory decline in AD patients and educational acquirements. Left: Linear and cubic curve-fitting on 582 observations projected on the progression space. The magnitude of the slope of the linear curve is 1.19 indicating a rapid memory decline as compared to cognitive. Right: Linear curve fitting for 200 most educated and 200 least educated patients.

3.4.5 Distribution of memory and cognitive decline

Figure 3.14 shows the distribution of projected dimensions (memory decline and cognitive decline) for each AD subtype after 24 and 48 months. In the progression space, along the positive direction of the y-axis, the memory decline increases, and along the negative direction of the x-axis, the cognitive decline increases. A low value on the x-axis indicates a higher cognitive decline, whereas a high value on the y-axis indicates higher memory decline. High progression rate has the highest memory and cognitive decline, which goes on reducing with a reduction in progression rate.

3.4.6 Selective features and AD progression rates

Figure 3.15 shows the distribution of the MMSE score after 6 and 12 months for each AD subtype at the 24^{th} month. Reduction in the MMSE score with increased progression is observed. A similar trend is observed in the distribution of the MMSE score for each AD subtype at the 48^{th} month. Further, there is an increase in functional assessment questionnaire (FAQ) total score with increasing progression rate for the 24^{th} and 48^{th} month AD subtypes as shown in Figure 3.15.

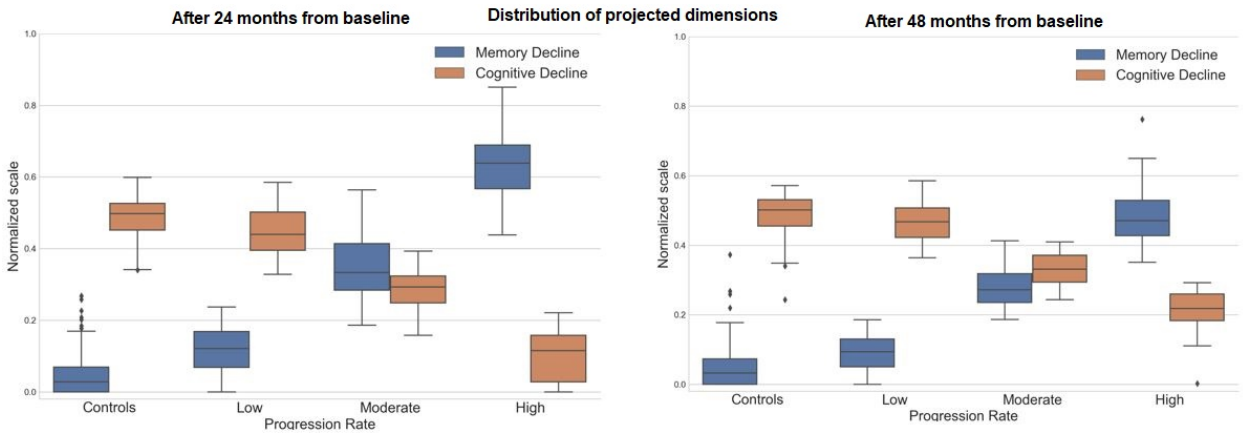

Figure 3.14: Distribution of projected dimensions (cognitive and memory decline) for each AD progression subtype. A lower numeric value on the cognition axis indicates high cognitive decline whereas a higher numeric value on the memory axis indicates high memory decline. A similar relationship can be observed in the figures. Left: after 24 months from baseline. Right: after 48 months from baseline.

3.4.7 Limitations and future directions

In this work, we discussed the share of different APOEε4 genotype for each progression rate and its correlation with cognitive performance. Future work should involve examining a few other genes that also have been closely related to the progression of AD for studying their interactions [120, 121]. As stated earlier in the paper, AD risk is associated with APOEε4 gene variants [117]. However, the progression space was constructed using only the time-variant clinical data. Therefore, APOEε4 data were not considered during the construction of the projection space. Moreover, the diagnosis of participants (control, MCI, AD) in the ADNI study is based on their clinical examinations, having a sensitivity of 70.9-87.3% as compared to the neuropathologic assessments, which are considered the gold standard for AD identification [122]. Hence, the discussed progression models suffer from the implicit noise involved in the diagnosis of the study participants. For future analysis, involving diagnosis with neuropathologic examinations may help scale down the ambiguity involved in the true status of participants [121].

The present analysis can be continued in various directions. Since both AD and PD are neurological diseases with AD primarily affecting memory advancing to influence motor functions and PD impacting movement and coordination progressing to hinder memory and other cognitive processes, exploring to project ADNI with a PD dataset might explain features responsible for PD progressing to AD or vice versa. Moreover, this study involved

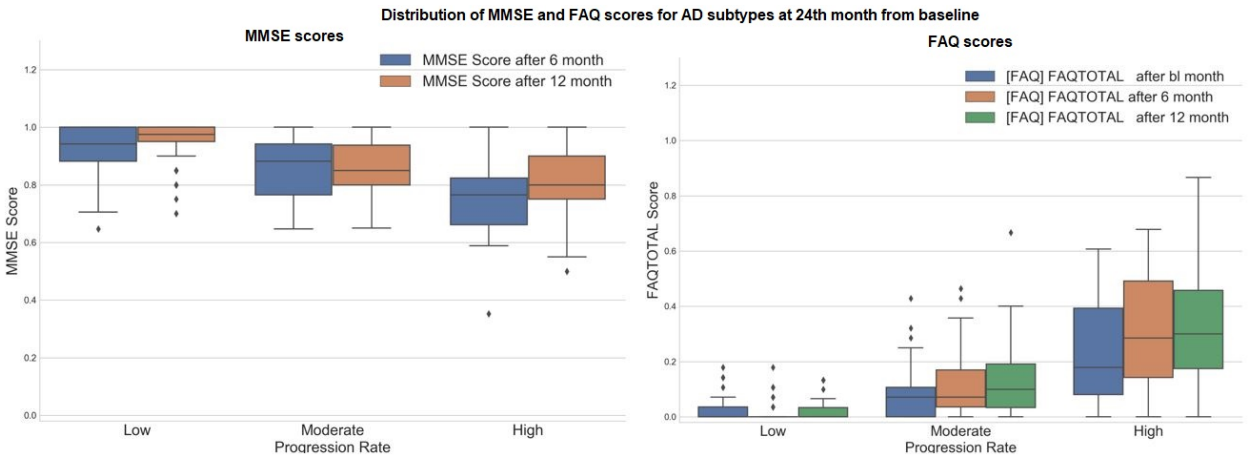

Figure 3.15: Distribution of the MMSE score for each AD subtype (subtypes at 24^{th} month). MMSE score decreases with an increase in progression rate. Right: Distribution of FAQ total score for each AD subtype (subtypes at 24^{th} month). FAQ total score increases with an increase in progression rate.

investigations only considering the clinical data for exploring AD progressions. Integrating further information such as neuroimaging or biomarker data may augment additional information in our analysis. Since we relied on the CV to gauge the performance of our models, validating the study on separate AD datasets such as AMP-AD [123] when it becomes available. Finally, the discussed analysis only focused on predicting progression space in AD. The proposed framework can be further adapted to study additional kinds of dementia, such as frontotemporal degenerations, Lewy body dementia, multi-infarct dementia, etc.

3.5 SUMMARY

This work clusters participants in distinct progression stages of AD and discusses an approach to predict the future rate of progression after the 24^{th} and 48^{th} months from baseline using longitudinal clinical data. Predicting disease progression serves as a paramount challenge in the therapy and cure of several elaborate diseases. This study is a step forward towards designing sophisticated machine-learning paradigms to facilitate early diagnosis of AD progression. Predicting AD progression rates would lead to better patient-specific attention by recognizing at an early stage the patients with a swift rate of progression. The proposed disease progression and trajectory prediction algorithms can help doctors and practitioners develop a methodical and organized course for clinical tests, which can be much more concise and effective in detection. These adaptations and modifications in clinics may help to diminish treatment and therapy costs for dementia. Further, no healing treatments

in AD modification exist. Hence, the capability to anticipate the trajectory of impending AD progression at the early stages of the disease is an advancement towards uncovering novel treatments for AD modification. The proposed analysis provides insights to inhibit or decelerate the progression of AD-related symptoms and subsequent deterioration in the characteristics of life that are accompanied by the disease.

CHAPTER 4: LEARNING AMYOTROPHIC LATERAL SCLEROSIS (ALS)

In this chapter, we review the work on *Identification and prediction of ALS subgroups through unsupervised and semi-supervised machine learning.* Similar to Alzheimer's and Parkinson's chapters, we use an unsupervised learning approach to identify new subtypes. However, unlike Parkinson's and Alzheimer's, which have a long multi-year course, we do not predict progression. Since ALS is a rapidly progressive disorder, we focus on subtype identification and survival analysis. We also use semi-supervised learning for enhancing subtype identification. A subset of this work was published previously as [52, 53, 54].

Amyotrophic lateral sclerosis is not a single entity, but rather represents a collection of syndromes in which the motor neurons degenerate. Synergistic with these multiple genetic etiologies, there is a broad variability in the clinical manifestations of the disease in terms of ages of symptom onset, site of onset, rate and pattern of progression, and cognitive involvement. Due to these variable presentations, counseling of patients about their individual risks and prognosis is limited. There is an unmet need for predictive tests that facilitate early detection and characterization of distinct disease subtypes as well as improved, individualized predictions of the disease course. The emergence of machine learning to detect hidden patterns in complex, multi-dimensional datasets provides unparalleled opportunities to address this critical need.

We used unsupervised and semi-supervised machine learning techniques to explore the clinical patterns of disease within a deeply-phenotyped, population-based collection of ALS patients. Our analysis distinguished distinct disease subtypes with a highly predictable site of onset. We have demonstrated robust replication of these findings in the independent validation cohort.

These data-driven results enable clinicians to deconstruct the heterogeneity within their patient cohorts. This knowledge could have immediate implications for clinical trials by improving the detection of significant clinical outcomes that might have been masked by cohort heterogeneity. We anticipate that machine learning models will improve patient counseling, clinical trial design, allocation of healthcare resources, and ultimately individualized clinical care.

4.1 INTRODUCTION

Although widely considered to be rare, amyotrophic lateral sclerosis (ALS, OMIM #105400) is one of the most common forms of neurodegeneration in the general population, accounting for approximately 6,000 deaths in the United States and 11,000 deaths in Europe annually [124, 125]. Characterized by progressive paralysis of limb and bulbar musculature, it typically leads to death within three to five years of symptom onset. Medications only minimally slow the rate of progression, and, as a consequence, treatment is primarily focused on symptomatic management.

Genetic advancements have shown that ALS is not a single entity, but rather represents a collection of syndromes in which the motor neurons degenerate [126]. Synergistic with these multiple genetic etiologies, there is a broad variability in the clinical manifestations of the disease in terms of ages of symptom onset, site of onset, rate and pattern of progression, and cognitive involvement. This clinical heterogeneity has hampered efforts to understand the cellular mechanisms underlying this fatal neurodegenerative syndrome, and has hindered clinical trial efforts to find effective therapies [127].

Given the importance of this clinical heterogeneity, it is not surprising that there have been myriad efforts to develop classification systems for ALS patients over the years. Examples include categorization based on family status [128], categorization based on clinical milestones [129], based on neurophysiological measurements [130, 131] and categorization based on the certainty of diagnosis [132]. The ability to identify the true number and type of subgroups with the ALS population and the capacity to reliably assign patients to these distinct subgroups would be a major step forward for the field. Though broadly useful, each of these existing classification systems suffer from a central problem, namely that it is unclear if they identify meaningful subgroups within the ALS population, or merely represent human constructs applied to the data based on preconceived notions.

Here, we applied novel unsupervised and semi-unsupervised machine learning techniques to explore the clinical patterns of disease within a deeply-phenotyped, population-based collection of ALS patients. With the unsupervised machine learning approach, our goal was to determine what subtypes of the disease might exist within this patient population and to see if we could reliably predict which subgroup an ALS patient might belong to. This process produced a multi-dimensional space that captures the topological and relative relationships of the ALS subtypes, mapping similar cases close together. The key advantage of

this machine learning approach is the ability to identify complex relationships in a uniquely unbiased and data-driven manner that moves beyond the traditional univariate approach.

Identified the ALS subtypes with the unsupervised learning approach, we then utilize the semi-supervised learning approach. We incorporated the physician's assessment of patients for the predictive task. With the trained model, we dissected the machine learning model to realize the underlying ALS structure which machine has learned. This representation is a close proxy to what the model "thinks" about data for it to classify patients as observed by the physician a year into the disease. Resulting in a more fine-tuned representation of ALS subtypes compared to the fully unsupervised.

Following the successful presentation of ALS subtypes, we built baseline predictor models using a supervised machine learning approach. These baseline predictors accurately predicted an individual patient's age of symptom onset and prognosis, a key concern whenever an individual is diagnosed with ALS. Such algorithms could provide a standardized approach to disease classification that minimizes inter- and intra-rater variability, which hampers multi-center clinical trial efforts.

4.2 METHODS

4.2.1 Study participants

The Piemonte and Valle d'Aosta Register for Amyotrophic Lateral Sclerosis (PARALS) [133] was used as the discovery cohort. PARALS included 2,858 patients who had been diagnosed with ALS according to El Escorial criteria [132], and represented incident cases who had been enrolled in Piedmont and Valle d'Aosta Registry for ALS between January 1, 1995, and December 31, 2016. Established in 1991, this population-based registry prospectively ascertains ALS cases within two regions of Northwestern Italy that include Turin and has a catchment population of nearly 4.5 million inhabitants [133]. Patients are followed longitudinally over the course of their illness, allowing the course of their illness to be observed. Informed written consent was obtained from all participants, and the study was approved by the ethics committee of A.O.U. City of Health and Science of Turin.

The replication cohort consisted of 1,097 patients who had been diagnosed with ALS according to the El Escorial criteria and represented incident cases who had been enrolled in the Emilia Romagna Region (ERR) ALS registry between January 1, 2009, and March

1, 2018. This registry prospectively collects demographic and clinical data on ALS cases incident within a Northwestern region of Italy that includes Modena and Bologna [134]. Similar to PARALS, the catchment population is 4.4 million, and patients are followed in a longitudinal manner. The methods for collecting clinical and demographic data are standardized across the PARALS and the ERR registries to facilitate comparisons between these two large epidemiological efforts. Informed written consent was obtained from all participants, and the study was approved by the ethics committees of the coordinating center and of the nine provinces of ERR.

4.2.2 Clinical data

For each cohort, a comprehensive and shared set of collected common data elements were selected for analysis. Table 4.1 provides a list of all clinical features as well as the analytical method they were used in.

Table 4.1: List of clinical parameters used in each stage of analysis.

Begin of Table 4.1					
#	**Clinical feature**	**Used in unsupervised and semi-supervised clustering**	**Used in post clustering analysis**	**Used in supervised age of symptom onset prediction**	**Used in supervised survival time prediction**
1	Cancer	Yes		Yes	Yes
2	Cancer type	No, due to irrelevance		No	No
3	Hypertension	Yes		Yes	Yes
4	Hyperthyroid	Yes		Yes	Yes
5	Hypothyroid	Yes		Yes	Yes
6	Diabetes	Yes		Yes	Yes
7	COPD	Yes		Yes	Yes
8	Smoker	No, due to high missingness	Yes	Yes	Yes
9	Parkinsonism	Yes		No	Yes
10	Corea	Yes		No	Yes
11	Ataxia	Yes		No	Yes
12	Marital status	Yes		Yes	Yes

Table 4.1 (cont.)

#	Clinical feature	Used in unsupervised and semi-supervised clustering	Used in post clustering analysis	Used in supervised age of symptom onset prediction	Used in supervised survival time prediction
13	Education	Yes		Yes	Yes
14	Cognitive status1	No, due to high missingness		No, due to high missingness	No, due to high missingness
15	Cognitive status2	No, due to high missingness		No, due to high missingness	No, due to high missingness
16	Cognitive impairment present	No, due to high missingness		No, due to high missingness	No, due to high missingness
17	Initial Dx was PLS	No, due to potential bias	Yes	No	No
18	elEscorial at Dx	Yes		Yes	Yes
19	elEscorial2	No, due to irrelevance		No, due to irrelevance	No, due to irrelevance
20	elEscorial3	No, due to irrelevance		No, due to irrelevance	No, due to irrelevance
21	FVC percent at Dx	Yes		Yes	Yes
22	Height	Yes		Yes	Yes
23	Weight at Dx	Yes		Yes	Yes
24	BMI at Dx	Yes		Yes	Yes
25	Weight 2 years prior to illness	Yes		No	Yes
26	BMI 2 years prior to illness	Yes		No	Yes
27	Rate of decline BMI per month	Yes		No	Yes
28	PEG inserted	No, due to possible data leakage	Yes	No	Yes
29	PEG days into illness	No, due to possible data leakage		No	Yes
30	BIPAP	No, due to possible data leakage	Yes	No	Yes

Table 4.1 (cont.)					
#	**Clinical feature**	**Used in unsupervised and semi-supervised clustering**	**Used in post clustering analysis**	**Used in supervised age of symptom onset prediction**	**Used in supervised survival time prediction**
31	BIPAP days into illness	No, due to possible data leakage		No	Yes
32	Tracheostomy	No, due to possible data leakage	Yes	No	Yes
33	Trach days into illness	No, due to possible data leakage		No	Yes
34	Place of birth	No	Yes	No	No
35	Place of residence	No	Yes	No	No
36	Family history of ALS	Yes		Yes	Yes
37	c9orf72 status	No, due to possible data leakage	Yes	Yes	Yes
38	Mutation present	No, due to potential bias	Yes	Yes	Yes
39	Mutated gene	No, due to potential bias	Yes	Yes	Yes
40	Mutation AA change	No, due to potential bias	Yes	Yes	Yes
41	Anatomical level at onset	Yes		No	Yes
42	Site of onset	Yes		No	Yes
43	Onset side	Yes		No	Yes
44	Clinical type at onset	No, due to high missingness	Yes	No	Yes
45	Clinical type at one year	No, due to data leakage	Yes	No	Yes
46	Sex	Yes		Yes	Yes
47	Age at onset	Yes		Predicting it	Yes
48	Age at diagnosis	Yes		No	Yes
49	Delay in Dx days	Yes		No	Yes

Table 4.1 (cont.)					
#	**Clinical feature**	**Used in unsupervised and semi-supervised clustering**	**Used in post clustering analysis**	**Used in supervised age of symptom onset prediction**	**Used in supervised survival time prediction**
50	Survival days	No, due to data leakage		No	Predicting it
51	Vital status	No, due to data leakage		No	No
52	ALSFRS1	Yes		Yes	Yes
53	ALSFRS2	Yes		Yes	Yes
54	ALSFRS3	Yes		Yes	Yes
55	ALSFRS4	Yes		Yes	Yes
56	ALSFRS5	Yes		Yes	Yes
57	ALSFRS6	Yes		Yes	Yes
58	ALSFRS7	Yes		Yes	Yes
59	ALSFRS8	Yes		Yes	Yes
60	ALSFRS9	Yes		Yes	Yes
61	ALSFRS10	Yes		Yes	Yes
62	ALSFRS11	Yes		Yes	Yes
63	ALSFRS12	Yes		Yes	Yes
64	First ALSFRS total	Yes		Yes	Yes
65	First ALSFRS days into illness	Yes		No	Yes
66	Rate of decline ALSFRS per month	Yes		No	Yes
End of Table 4.1					

4.2.3 Outcome measures

Each patient in the discovery and replication cohort was assigned to one of six clinical subtypes (bulbar ALS, respiratory ALS, flail arm ALS, classical ALS, pyramidal ALS, and flail leg ALS) based on the classification system published by Chiò and colleagues [129]. In contrast to other classification systems, these subtypes are determined based on clinical features of the patient one

year after symptom onset. We compared the subtype clusters defined by our unsupervised machine learning approach to the clinical subtype assigned by the Chiò classification system. For this reason, the clinical subtypes assigned by the Chiò classification system were not entered into the machine learning algorithm.

Age at symptom onset was defined as the age of the patient when they first developed symptoms of weakness in their bulbar, respiratory, or limb muscles. Survival time was defined as the number of days from symptom onset to (a) death; (b) tracheostomy and permanent ventilation; or (c) date of the last followup if the patient was still alive.

4.2.4 Filtering of clinical data

Clinical data were filtered prior to analysis. Features that are not meaningful for subtyping of the ALS patients (e.g., cancer type, place of birth) or that could introduce bias (e.g., tracheostomy, initial Dx was PLS) were omitted from the analysis. Samples with missing values (n = 497 in the discovery cohort, n = 108 in the replication cohort) in the revised ALS Functional Rating Score (ALSFRS-R) were also omitted, as this parameter was found to be a powerful predictor of subtype. As machine learning methods used in this paper only accept numeric input and are sensitive to input scaling, categorical features were *one-hot encoded*. *Min-max normalization* was also applied to numeric features to preserve the shape of the distribution and ensure they were within a zero to one range. Ultimately, 2,361 cases in the discovery cohort and 989 in the replication cohort pass the filtering step.

4.2.5 Data imputation

After filtering, no missingness was found in most features except in FVC percent at Dx, BMI 2 years prior to illness, rate of decline BMI per month, weight two years prior to illness, BMI at Dx, height, and weight at Dx. These features were missing at random with missingness of 15-20%. We used the *k-Nearest Neighbor* (kNN) imputation method with $k = 5$ neighbors to preserve the clusters [135].

4.2.6 Unsupervised subtype identification using Uniform Manifold Approximation and Projection

We used an unsupervised clustering approach to identify ALS subtypes. We applied *Uniform Manifold Approximation and Projection* (UMAP) to processed data. UMAP [136] is a machine learning approach that is primarily used for non-linear dimension reduction. UMAP produces a data embedding by searching for a low dimensional projection of the data that has the closest

possible equivalent fuzzy topological structure. This non-linear dimension reduction preserves the local and global structures existing within the data, along with reproducible and meaningful clusters.

4.2.7 Semi-supervised subtype identification using multilayer perceptron neural network

To enhance our ability to resolve the different types of spinal-onset ALS, we pre-processed the data using a *multilayer perceptron neural network* consisting of five hidden layers with 200, 100, 50, 25 and 3 neurons 4.1. This supervised technique reduced the 72 input clinical features of our dataset down to three dimensions. The dimension-reduction network was trained on the *'clinical type at one-year'* feature using a *Softmax* classifier. After training the network with ten-fold cross-validation, the data is then once more used as input for the forward loop of the network. This time, the last hidden layer activations, which represent a dimension reduction from 72 to 3 dimensions, were used as input for the next t-SNE algorithm to improve the separation of the clusters further.

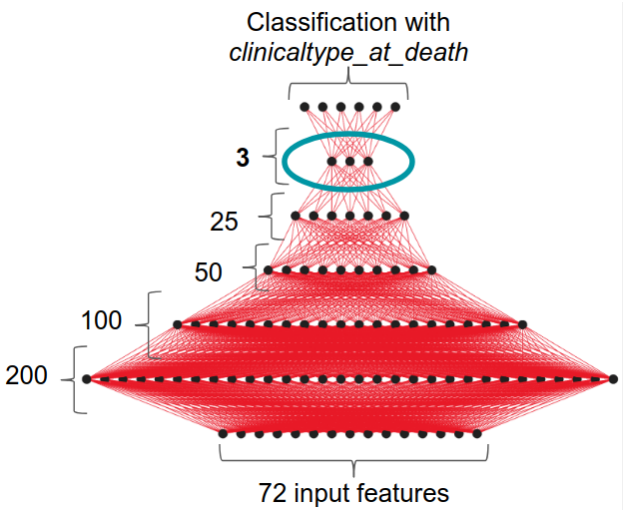

Figure 4.1: Architecture of the multilayer perceptron neural network (MLP) used for the semi-supervised subtype identification of ALS. The neural network consists of five hidden layers with 200, 100, 50, 25, and 3 neurons. After training the network with ten-fold cross-validation, the activations of the last hidden layer are used as input for the next algorithm, t-SNE. The network compresses the data to 3 dimensions, acting as a dimension reduction technique.

t-distributed Stochastic Neighbor Embedding (t-SNE) [137] is a machine learning approach that is primarily used for the visualization of high-dimensional data within a lower-dimensional manifold. It calculates the pairwise densities of the data points in the original and the reduced space and then minimizes the difference between both densities. This non-linear dimension reduction preserves the local and global structures existing within the data.

4.2.8 Supervised prediction of the age of symptom onset and survival time

We assessed the ability of three supervised machine learning algorithms (Random forest [91], LightGBM [138], and XGBoost [139]) to predict the age of symptom onset and to predict survival time. Random forest algorithms leverage decision trees by training a high number of them in parallel and then evaluating the combined result. Each tree is trained on a randomly sampled subset of the training data, and at each additional randomness is added by sampling from the possible features to split on. This added randomness through the data and feature bagging ensures variance in the different trees. The results of the trees are then averaged to get a final prediction.

LightGBM and XGBoost are gradient boosting algorithms. Contrary to the random forest, who builds a multitude of trees in parallel, gradient boosting works by combining learners in sequence. In the following, the used learners are decision trees. Decision trees that usually are not to their maximum depth developed, called weak learners, are trained on the residuals of their predecessors, thereby putting emphasis on the previous misclassification.

The validity of the approach was assessed by dividing our cohort between a training dataset and a test dataset. Hyperparameters were tuned on the training cohort using five-fold cross-validation. Four of those were used as a training set for optimization, while the fifth is used as a validation set to check the results. To tune the hyperparameters, we build a grid of possible parameter combinations and use a randomized grid search to find a good combination that maximized the validation accuracy. Finally, a model was trained using the complete training dataset and the optimal hyperparameter combination, and then evaluated on the test set.

SHAP (SHapley Additive exPlanations) approach [140]. SHAP is a unified approach to explain the output of any machine learning model by assigning each feature an importance value based on the Shapley value. Shapley values are used in game theory to determine the contribution of players to success.

4.3 RESULTS

4.3.1 Identification of ALS subtypes

We applied unsupervised and semi-supervised machine learning approaches to our large dataset consisting of 72 clinical features collected from 2,858 patients diagnosed with ALS obtained from a prospective, population-based registry over a ten-year period. To do this, we used an unsupervised UMAP algorithm alongside with semi-supervised approach of pre-processing the data with a multi-layered perceptron neural network and using the dimension-reduced output as the input for

the t-SNE algorithm. Both approaches were able to define distinct clusters of patients, representing subtypes of ALS. Figure 4.2 left column shows the result of the machine-learned topologies. These projections show the relative distance of each patient in terms of their disease characteristics. Patients have formed into multiple global and local clusters. By color-coding the patient's clinical type at one year according to the Chiò classification system, comparison showed that each of these clusters corresponded to one of the six clinical subtypes previously defined by the Chiò classification system (bulbar ALS, respiratory ALS, flail arm ALS, classical ALS, pyramidal ALS, and flail leg ALS).

In order to ensure the generalizability and validity of the results, we replicate the ALS subtype identification in the independent replication cohort. Figure 4.2 right column shows the identified subtypes in the independent replication cohort. We see that the identified subtypes in the replication cohort are similar to the ones in the discovery cohort. The discovery and replication cohorts are clinically different cohorts and recruited from different populations. The replication of our results in the replication cohort that was recruited with a different protocol shows the strength of our study's methodology. We demonstrate that if we ascertain the same phenotypes using standardized scales, we can reliably discern the same subtypes. This makes our results robust, generalizable, and the clinical subtypes reproducible.

While using the aforementioned unsupervised learning approaches, we were able to visualize, cluster, and interpret ALS clusters, interpretability of the reduced dimensions remain difficult. Most algorithms must make trade-offs, and our utilized methods are no exception. Most non-linear dimension reduction techniques, including t-SNE and UMAP, lacks the strong interpretability of reduced dimensions. On the other hand, linear transformations such as Principal Component Analysis (PCA) and Non-Negative Matrix Factorization (NMF) produce interpretable reduced dimensions but are unable to capture relationships between complex clinical phenomena accurately. In particular, the dimensions of the UMAP embedding space have no specific meaning, unlike PCA where the dimensions are the directions of greatest variance in the source data.

4.3.2 Prediction of survival time for patients diagnosed with ALS

Next, we applied machine learning approaches to predict survival time among ALS patients. The most accurate survival prediction was provided by XGBoost modeling that incorporated the clinical subtype identified by our earlier analysis (mean absolute error = 378.1 days). We determined the clinical features that influenced survival using the SHAP approach that is designed to identify complex relationships in multi-dimensional datasets (Figure 4.3). Although the various features interact with each other within the model in a compound manner, some simple observations can be made. The most influential feature on prognosis was the *rate of decline in ALSFRS score per*

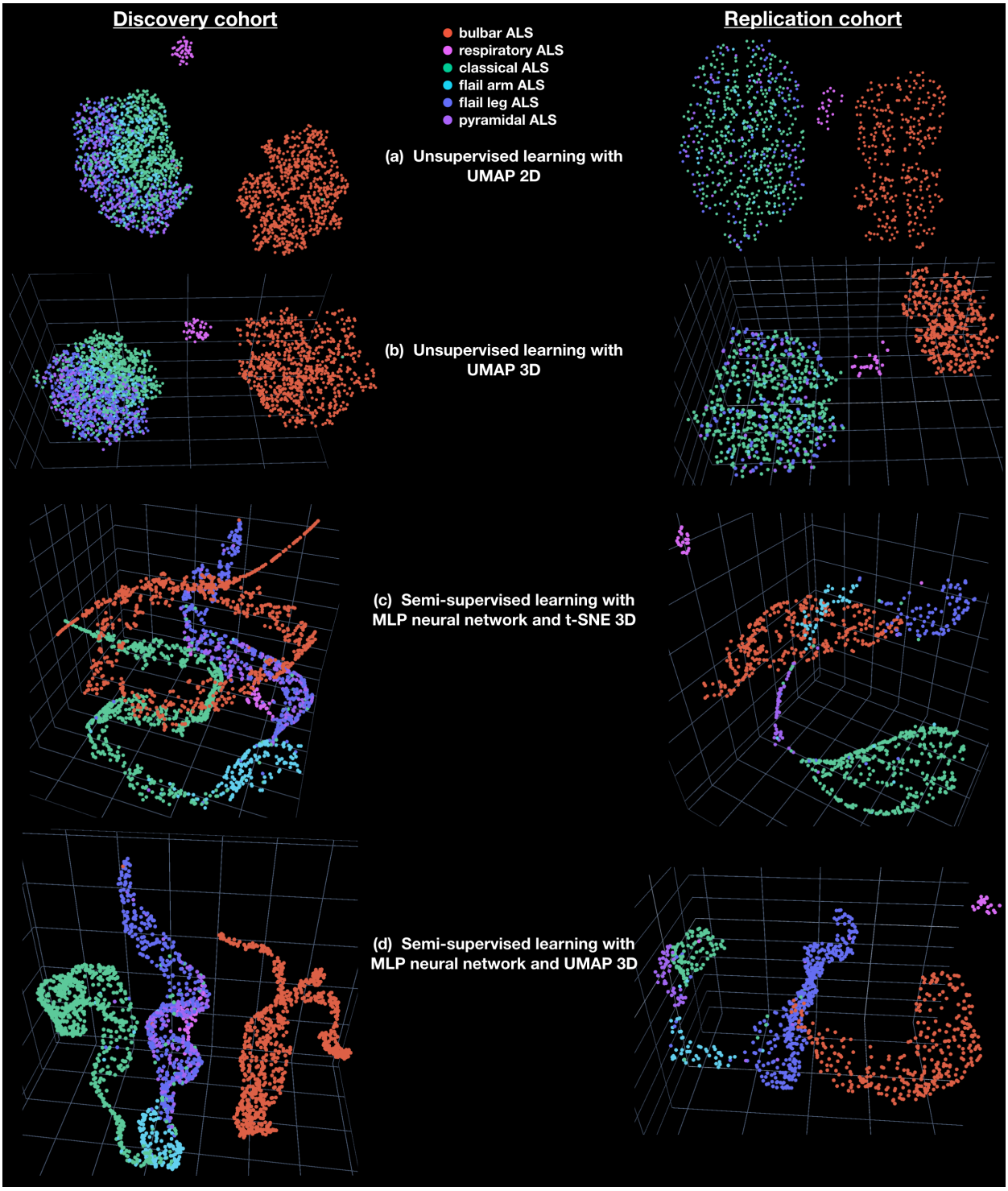

Figure 4.2: Identified ALS subtypes in the discovery and replication cohorts. Left column shows the projections of the discovery cohort and right column projections of the replication cohort. Each row corresponds to the utilized machine learning technique. Patients are color-coded with the clinical type at one year, according to the Chiò classification system. The comparison shows that each of these clusters corresponded to one of the six clinical subtypes previously defined by the Chiò classification system (bulbar ALS, respiratory ALS, flail arm ALS, classical ALS, pyramidal ALS, and flail leg ALS).

month, a widely accepted indicator of disease progression rate with higher values associated with poorer prognosis. The next most influential feature was the number of days after symptom onset that the ALSFRS scoring was first conducted. This variable acts as a surrogate for the delay in diagnosis, a well-established predictor of survival in ALS. The third most influential feature was the *forced vital capacity percentage at diagnosis*, a measure of remaining respiratory function with lower values predicting a shorter survival time.

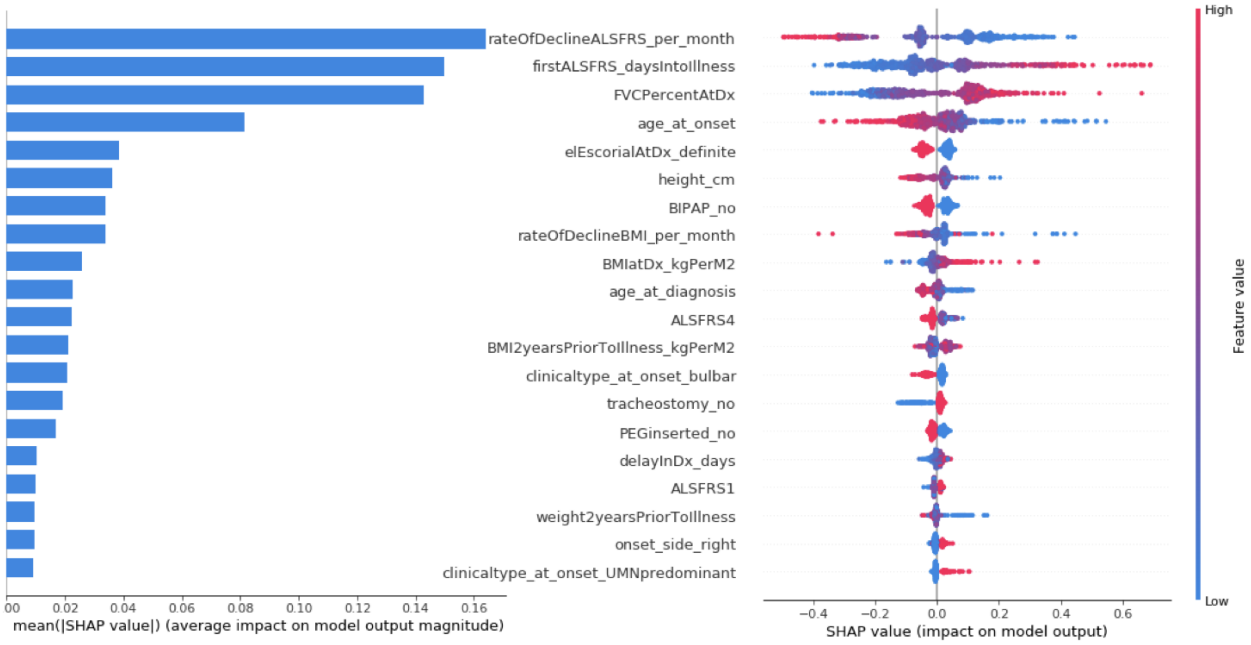

Figure 4.3: Clinical features influencing survival prediction. (Left) Bar plot showing the most influential clinical features based on SHAP values for the XGBoost prediction algorithm. (Right) Detailed view of the influence of clinical features on the XGBoost prediction algorithm.

4.3.3 Prediction of the age of symptom onset

When predicting the age at the onset of ALS, XGBoost was again the best performing algorithm (MAE = 7.7 years). The most important features of the prediction output were whether the patient has hypertension and the degree of education (Figure 4.4). The observation that education is associated with higher age at symptom onset is consistent with our recent publication describing the protective effect of education on the risk of developing ALS based on linkage disequilibrium regression score [53].

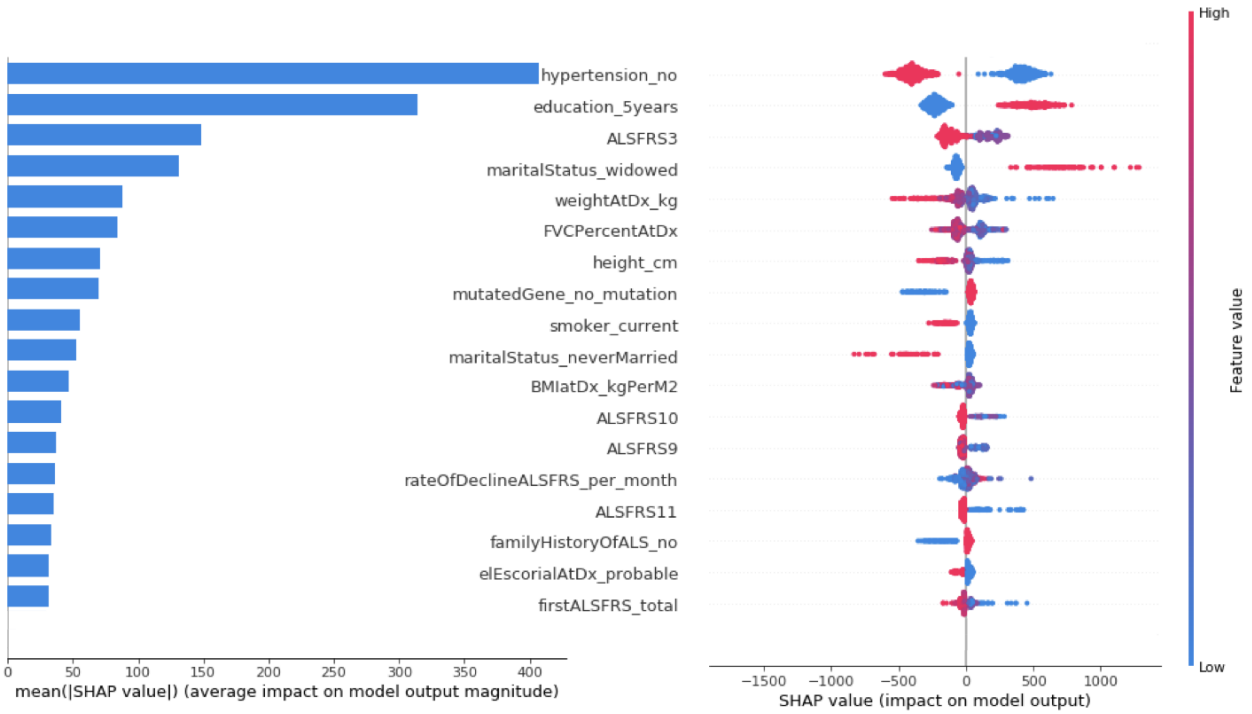

Figure 4.4: (Left) Barplot of the mean absolute SHAP values of the features for the XGBoost algorithm, showing the influence on the result for the biased case for predicting the age at onset. (Right) Detailed view of the influence of features on the XGBoost prediction on the unbiased data of the age at onset prediction.

4.4 DISCUSSION

There were two main obstacles to unraveling the clinical heterogeneity of ALS. First, there has been a paucity of large enough clinical databases. The PARALS is a large, population-based registry of ALS collected over a ten year period. Second, the clinical heterogeneity was clearly multidimensional and complex. Previous efforts based on univariate or incorporating a small number of variables have been spectacularly unsuccessful. Here, we benefit from the recent development of machine learning. Though it is a broad concept, the main benefit to us is that it reduces/deals with the complexity of the dataset.

Using an unbiased, data-driven approach, we identified six subtypes of ALS within our large discovery cohort and replicated these findings in an independent population-based cohort. These subtypes corresponded to those previously defined by the Chiò classification system, demonstrating the utility of that approach. The key difference of this classification system compared to others is that it relies on observing the patient one year after symptom onset. This lag in assigning categories allows the patient time to declare their category. Notably, our new algorithm provides the ability to predict the patient's category at an earlier time point.

It is notable that the machine learning algorithm arrived at the same conclusion as the Chiò and colleagues. We do not maintain that the machine learning algorithm is better at defining categories than this group of experienced ALS neurologist. Instead, we say that an unbiased, data-driven approach designed at identifying relationships within high dimensional data arrived at the same conclusion as Chiò and colleagues. Thus, we believe that our data provide prima facie evidence that the Chiò classification system is identifying the true and meaningful subgroups that exist within the ALS population.

4.5 SUMMARY

In this study, we addressed the complexities of ALS disorder. This work identifies the data-driven subtypes of ALS and discusses an approach to predict the disease course. Predicting disease subtypes serves as a paramount challenge in the therapy and cure of several elaborate diseases. This study is a step forward towards designing sophisticated machine-learning paradigms to facilitate early diagnosis of ALS. Predicting ALS subtypes would lead to better patient-specific attention by recognizing at an early stage the patients with a swift rate of progression. The proposed disease prediction algorithms can help doctors and practitioners develop a methodical and organized course for clinical tests, which can be much more concise and effective in detection. These adaptations and modifications in clinics may help to diminish treatment and therapy costs for ALS. Further, the capability to anticipate the subtypes of ALS at the early stages of the disease is an advancement towards uncovering novel treatments for ALS modification. The proposed analysis provides insights to inhibit or decelerate the progression of ALS-related symptoms and subsequent deterioration in the characteristics of life that are accompanied by the disease. In the future work, we will perform an in-depth analysis of identified subtypes and their genetic profiles.

CHAPTER 5: LEARNING INTENSIVE CARE UNIT (ICU)

In this chapter, we review the work on *Analysis and prediction of unplanned intensive care unit readmission*[1]. Unlike previous chapters where the health issue has a multi-year course, in the ICU, data is short-term. Short-term data such as sensory and monitoring device data has a high frequency of data collection but with a time window of hours to days. Due to the data collection frequency, different predictive machine learning methods are used to encapsulate the time-sensitivity of small fluctuations in the patient's status. This work was published previously as [55].

Unplanned readmission of a hospitalized patient is an indicator of patients' exposure to risk and an avoidable waste of medical resources. In addition to hospital readmission, intensive care unit (ICU) readmission brings further financial risk, along with morbidity and mortality risks. Identification of high-risk patients who are likely to be readmitted can provide significant benefits for both patients and medical providers. The emergence of machine learning solutions to detect hidden patterns in complex, multi-dimensional datasets provides unparalleled opportunities for developing an efficient discharge decision-making support system for physicians and ICU specialists.

We used supervised machine learning approaches for ICU readmission prediction. We used machine learning methods on comprehensive, longitudinal clinical data from the MIMIC-III to predict the ICU readmission of patients within 30 days of their discharge. We incorporate multiple types of features including chart events, demographic, and ICD-9 embeddings. We have utilized recent machine learning techniques such as Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM), by this we have been able to incorporate the multivariate features of EHRs and capture sudden fluctuations in chart event features (e.g. glucose and heart rate). We show that our LSTM-based solution can better capture high volatility and unstable status in ICU patients, an important factor in ICU readmission. Our machine learning models identify ICU readmissions at a higher sensitivity rate of 0.742 (95% CI, 0.718-0.766) and an improved Area Under the Curve of 0.791 (95% CI, 0.782-0.800) compared with traditional methods. We perform in-depth deep learning performance analysis, as well as the analysis of each feature contribution to the predictive model.

This work highlights the ability of machine learning models to improve our ICU decision-making accuracy and is a real-world example of precision medicine in hospitals. These data-driven solutions hold the potential for substantial clinical impact by augmenting clinical decision-making for physicians and ICU specialists. We anticipate that machine learning models will improve patient counseling, hospital administration, allocation of healthcare resources and ultimately individualized clinical care.

[1]This research was assisted by Yu-Wei Lin and Yuqian Zhou as documented in their thesis.

5.1 INTRODUCTION

Unplanned hospital readmission is an indicator of patients' exposure to risk and an avoidable waste of medical resources. To address the unplanned readmission issue, in 2010, the Affordable Care Act (ACA) created the Hospital Readmissions Reduction Program to penalize the hospitals whose 30-day readmission rates are higher than expected [141]. According to data released by the Centers for Medicare & Medicaid Services (CMS), since the program began on Oct. 1, 2012, hospitals have experienced nearly $2.5 billion of penalties assessed on hospitals for readmissions, including an estimated $564 million in fiscal year 2018, $144 million more than in 2016 [142].

In addition to hospital readmission, intensive care unit (ICU) readmission brings further financial risk, along with morbidity and mortality risks [143, 144]. Premature ICU discharge may potentially expose patients to the risks of unsuitable treatment, which further leads to an avoidable mortality [145]. Reportedly, the mortality rates of ICU readmitted patients range approximately from 26% to 58% [146, 147, 148]. Surprisingly, even in developed countries, hospitals suffer from high ICU readmission rates, around 10% of patients will be readmitted back to ICU within a hospital stay [143]. Moreover, there is an escalating trend in the U.S. for ICU readmission rates rising from 4.6% in 1989 to 6.4% in 2003 [144]. Thus, making ICU readmission rates one of the critical quality indicators in the performance evaluation of ICU.

To reduce avoidable ICU readmission, hospitals need to identify patients with a higher risk of ICU readmission [149]. Identified patients will stay longer in the ICU and will not be exposed to readmission risks. Moreover, the additional medical resources that would have been used in unnecessary readmission can be reallocated more efficiently considering the scarcity of ICU resources compared to the general hospital. Ultimately, an efficient decision-making support system can have significant impact by assisting hospitals and ICU physicians identifying patients with high readmission probability. We can use machine learning and artificial intelligence techniques to build such decision-making support systems. Data-driven predictive models aimed at predicting ICU readmission may be built using various datasets including administrative claims [150, 151, 152], insurance claims, and electronic health records (EHRs). Among these datasets, insurance claims are not suitable for real-time prediction [153] electronic health records (EHR) have shown to provide appropriate data for medical decision-making support solutions. A systematic review of readmission prediction models [154], summarizes 26 unique readmission prediction models of which 23 models rely on EHR including the most recent work on predicting all-cause 30-day readmission by Jamei et al. [153] which proposed an accurate and real-time prediction model based on neural networks.

Even though multiple studies have developed predictive models to tackle the problem of identifying patients with a high risk of readmission, we are still far from a comprehensive practical solution.

Overall, these studies have five main drawbacks. First, the scope of some predictive models is limited to a specific disease or treatment rather than a general solution. For instance solutions were focused on heart failure [155], HIV [156], diabetes [157], and kidney transplants [158]. Second, no model has been able to predict ICU readmissions to a satisfactory degree yet [159]; most models suffer from a low sensitivity of around 0.6 to 0.65 [145, 153, 159]. Third, most models do not utilize the sequential data structure and time series feature of many EHR parameters which can lead to information loss [160]. Last, very few attempts to understand and interpret the predictive model. Feature interpretation, as well as decision making logic, reliability, and robustness analysis of the machine learning models is crucial, and more imperative for clinical applications. This task is much more complex for deep learning techniques, which has made recent works short of explaining the decision making logic and model interpretation [161, 162].

In this study, we focus on the analysis and prediction of unplanned ICU readmission using recent deep learning techniques and utilizing time series feature of data. We propose a recurrent neural network (RNN) architecture with long short-term memory (LSTM) layers to enhance the predictive model by incorporating the time series data. We also incorporate low-dimensional representations (also called embeddings) of medical concepts (e.g. diseases ICD-9 code, treatment procedure, and laboratory tests) as the input of our model [150, 163]. Finally, we test, validate, and explain the proposed methods using the MIMIC-III dataset [164], containing more than 40,000 patients' information and 60,000 ICU admission records, over a 10 year period [164]. We leverage this extensive dataset to develop predictive model which provides clinicians with the much needed decision-making support. This data-driven approach can help prevent the inappropriate discharge or transfer of patients at high-risk of ICU readmission along with reducing the associated costs and penalties.

5.2 METHODS

To accompany this report, and to allow independent replication and extension of our work, we have made the code publicly available under GPLv3 for use by non-profit academic researchers at `https://github.com/Jeffreylin0925/MIMIC-III_ICU_Readmission_Analysis`. The code is part of the supplemental information; it includes the step-by-step instructions of the statistical and machine learning analysis.

5.2.1 Dataset construction

The readmission dataset is constructed from the MIMIC-III Critical Care Database. MIMIC-III consists of the health-related EHR data of more than 40,000 patients in the Intensive Care Units (ICU) of the Beth Israel Deaconess Medical Center between 2001 and 2012. One patient may have

multiple in-hospital records in the dataset. Following the data screening process stated in [157], we first screen out the patients under age 18 and remove the patients who died in the ICU. This results in total number of 35,334 patients with 48,393 ICU stays. We then split the processed patients into training (80%), validation (10%), and testing (10%) partitions to train our model and conduct a five-fold cross-validation. Note that one patient may have multiple records, so the number of items may not equal in each fold.

To construct the dataset for ICU readmission, we categorize all selected patients and their corresponding ICU stays records into positive or negative cases. Specifically, the following cases are considered to be positive patient stays:

- 3,555 records: the patients were transferred to low-level wards from ICU, but returned to ICU again,
- 1,974 records: the patients were transferred to low-level wards from ICU, and died later,
- 3,205 records: the patients were discharged, but returned to the ICU within the next 30 days,
- 2,556 records: the patients were discharged and died within the next 30 days.

Positive cases are regarded as the ones where the patients could benefit from a prediction of readmission before being transferred or discharged. Negative cases, on the contrast, are those where the patient does not need ICU readmission. Specifically, patients who were transferred or discharged from ICU and did not return and are still alive within the next 30 days are considered to be negative cases.

5.2.2 Feature extraction

In this section, we introduce the features and the time series window we use for the ICU readmission prediction task. For temporal information modeling of the time series ICU records, we use the last 48-hour data of each ICU stay. The last 48 hours before the patient is discharged or transferred are found to be the most informative data for prediction of readmission [165, 166]. To cope with the problem of data missingness, we use Last-Observation-Carried-Forward (LOCF) imputation method. In cases where the last hour is missing, we include an indicator for missingness.

We use three categories of features for developing our readmission prediction model, namely chart events, ICD-9 embeddings, and demographic information of the patients. First, chart events category, which are extracted from health care provider (e.g., physicians and nurses) notes. Chart events represent the patient's' physiological conditions based on the experts' observation and opinions [159]. Second, patient variables like chronic diseases. This category has been found to strongly

Table 5.1: 17 Types of features in the chart events.

Chart Events	**Dim**	Normal
1. Glasgow coma scale eye opening	8	4 Spontaneously
2. Glasgow coma scale verbal response	12	5 Oriented
3. Glasgow coma scale motor response	12	6 Obeys Commands
4. Glasgow coma scale total	13	15
5. Capillary refill rate	2	Normal <3 secs
6. Diastolic blood pressure	1	70
7. Systolic blood pressure	1	105
8. Mean blood pressure	1	87.5
9. Heart Rate	1	80
10. Glucose	1	85
11. Fraction inspired oxygen	1	0.21
12. Oxygen saturation	1	97.5
13. Respiratory rate	1	15
14. Body Temperature	1	37
15. pH	1	7.4
16. Weight	1	80.7
17. Height	1	168.8

associate with ICU readmission risk [145, 165]. Third, basic demographic information, such as gender, age, race. This category has also been demonstrated as important factors in the readmission prediction [153]. In this study, we leverage all of the above-mentioned feature categories and their time series information for the readmission prediction task. We also extract both basic and advanced statistical features from the chart events in order to compare our proposed model to traditional methods as baseline such as logistic regression.

Chart events: We extract 17 types of time series features from chart events within a 48-hour window. The raw features include both numerical (e.g., diastolic blood pressure) and categorical items (e.g., capillary refill rate). Details of these 17 features and their dimensions are shown in Table 5.1, along with their normal median value in the humans. We use the normal values later in the discussion section for machine learning model interpretation. In total 59 dimensions are constructed from the chart events; the increased number is due to the one-hot encoding of the categorical features. To identify and overcome the missing records in the chart events, we create a 17-dim binary indicator feature, appended to the chart events feature. This feature indicates whether the record for each type of chart event exists.

ICD-9 embeddings: Chronic diseases are found as one of the most important factors associated with later readmissions [165]. However, this information tends to be sparse in an EHR dataset, making them one of the most challenging to analyze with machine learning methods. In order to

Table 5.2: Demographic Features.

Chart Events	**Dimension**	**Option**
1. Gender	2	Male/Female
2. Age	1	18-120
3. Insurance Type	5	Government, Self, Medicare, Private, Medicaid
4. Race	6	Asian, Black, Hispanic, White, Other, No Information

address the data sparsity of disease information in the EHR, we apply the approach presented in [150] to compute a pre-trained 300-dimension embedding for each ICD-9 code recorded. Utilizing a lower dimension embedding of ICD-9 benefits the model training process by avoiding a sparse representation and applying the relationship information among different diseases. For a patient with multiple diseases, we simply take the addition of embeddings of all the diseases in order to construct the feature.

Demographic features: The demographic features consist of the patient's gender, age, race, and insurance type. Details of this category and its corresponding dimensions are summarized in Table 5.2. We include the insurance type as it could potentially influence the discharge/transfer rate. For example, although unlikely, an insurance type uninsured could lead to insufficient payment and might result in an unexpected discharge. In total there are 14 dimensions for the demographic category.

Statistical features for baseline models: For the purpose of comparison to the traditional methods, we also extract the statistical features within each 48-hour window. We include the slopes and intercepts of the regression line (a and b in $y = ax + b$) as separate features to characterize the linear trend for continuous data including the numerical chart events. Linear regression approach has been widely used in ICU readmission prediction [167, 168, 169]. For the categorical data such as capillary refill rate, we follow the approach in [170, 171] to extract the mean and majority value over the total time period after transforming categorical events into binary or ordinal. Figure 5.1 shows an example of extracted statistical features for the baseline model comparison. After computing the statistical features, each 48-hour data window will become one single data point, resulting in 71 dimensions of chart events.

Furthermore, in order to include chart events' volatility, we include more complex statistical features to enhance the regression model for better baseline model comparison. For numerical data, we extract: (i) quadratic term, (ii) standard deviation, (iii) mean absolute deviation, and (iv) R^2. Adding these statistical features, results in the increase of dimensions from 2 to 6 for numerical features. For categorical data, we extract: (i) majority value, and (ii) how often the value switches. These statistical features enable us to better capture the volatile nature of ICU

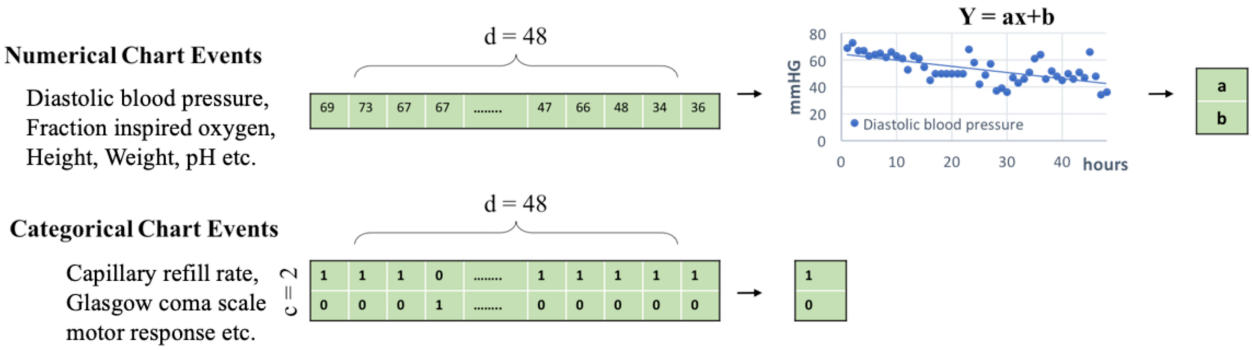

Figure 5.1: Statistical feature computation. For numerical chart events, we conduct linear regression on the 48-hour data points and record the rate and bias value as the feature. For categorical events, we simply compute the average occurrence of the categories.

events in the traditional baseline models. We call the earlier statistics "basic statistical features (B-STAT)" and the combination of basic and more complex statistical features "advanced statistical features (A_STAT)" for the rest of this paper.

5.2.3 Machine learning model structure

Baseline models: The first baseline model that we include is the logistic regression models. In this study, we implement logistic regression with both L1 and L2 regularization penalty. We further train three conventional machine learning models as our baseline, including Naive Bayes, Random Forest, and Support Vector Machines (SVM).

Convolutional neural network (CNN) model: We also implement a CNN-based model for comparison to our LSTM model. CNN-based models are found useful in analyzing longitudinal EHR data [161]. Shown in Figure 5.2, we use a multi-filter CNN structure introduced in [172]. We use the CNN model on a comprehensive and longitudinal representation of data with 18,720 dimensions as shown in Figure 5.3. We conduct the convolution on the time axis with 48-hour time window and D dimension using filter size 2, 3 or 4 accordingly. The computed feature maps are finally concatenated and fully connected to a dense decision layer with one output neuron.

Long short-term memory (LSTM) model: LSTM networks are found well-suited to making predictions based on time series data, especially for clinical measurements where there can be lags of unknown duration and missing values in a time series [173]. Figure 5.4 shows our utilized LSTM model. We use a bidirectional LSTM combined with an additional LSTM layer, followed by a dense decision layer with one output neuron activated by a sigmoid function. Overall, we have 16 hidden units in our LSTM layer. Bidirectional LSTM learns the temporal information across the whole training window. Considering an ICU stay record with a length of 48 hours, observation at each

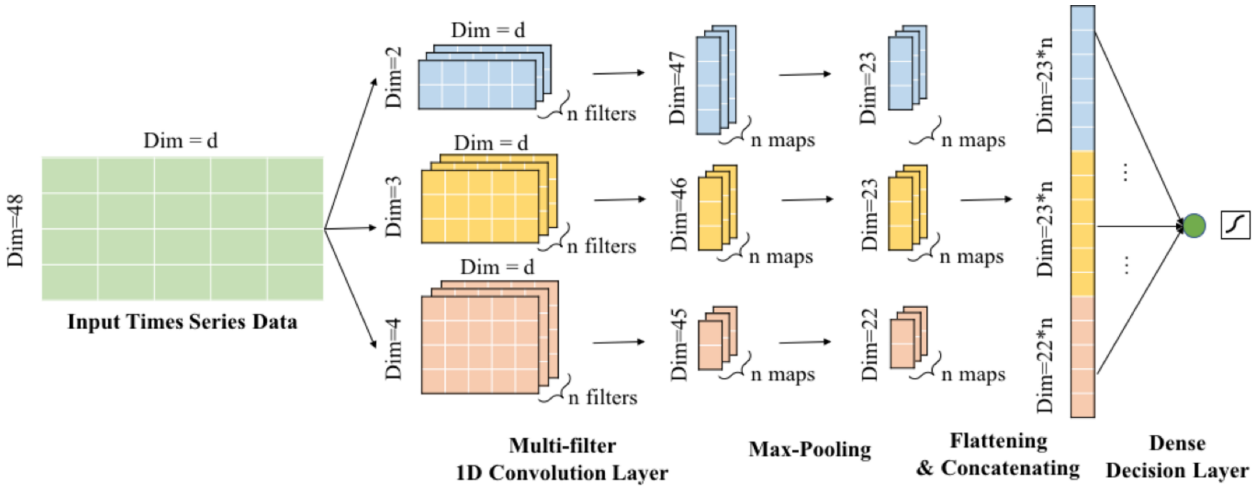

Figure 5.2: The 1D multi-filter convolutional neural network. We conduct the convolution on the time axis with 48-hour time window and D dimension using filter size 2, 3 or 4 accordingly. The computed feature maps are finally concatenated and fully connected to a dense decision layer with one output neuron.

hour is denoted by $x_t \in \mathbb{R}^{1\times D}$, where t is an integer from 1 to 48, and D is the feature dimension size. The output of a single LSTM cell can be computed by the following equations,

$$\begin{aligned} i_t &= \sigma(W_i.[h_{t-1}, x_t] + b_i) \\ f_t &= \sigma(W_f.[h_{t-1}, x_t] + b_f) \\ \hat{C}_t &= \tanh(W_c.[h_{t-1}, x_t] + b_c) \end{aligned} \tag{5.1}$$

$$\begin{aligned} C_t &= f_t \odot C_{t-1} + i_t \odot \hat{C}_t \\ o_t &= \sigma(W_o.[h_{t-1}, x_t] + b_o) \\ h_t &= o_t \odot \tanh(C_t) \end{aligned}$$

The above functions can be simply denoted by $h_t = LSTM(h_{t-1}; x_t)$. We utilize the hidden value of the last time stamp to predict the readmission possibility, thus the final output after going through the dense layer would be,

$$r_T = \sigma(W_r.h_{48} + b_r) \tag{5.2}$$

where σ is the indicator of the sigmoid activation function, and the r_T represents the prediction probability of whether this patient with the ICU stay record will be readmitted, ranging from zero to one. The dimension of h_t is $\mathbb{R}^{1\times 16}$, therefore the $W_r \in \mathbb{R}^{16\times 1}$. We also use binary cross entropy loss to update the weights. In addition to separate CNN and LSTM based models, we also implement and compare the performance of the LSTM and CNN combination models. We implemented all the models using Keras based on the benchmark code of [160]. The learning rate of training

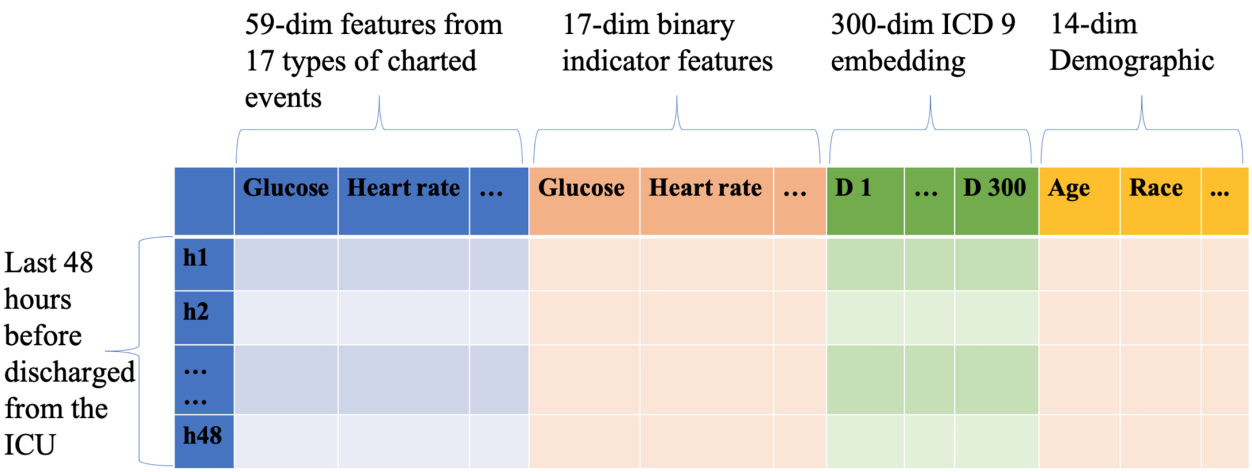

Figure 5.3: The data structure of input data used with CNN and LSTM models. D: dimension, h: hour.

was set to 1e-3, and we used Adam optimizer to train the model with beta 0.9. We trained at most 50 epochs and selected the model with the highest AUC on the validation partition following the logic in [173]. During the evaluation, we set up the decision threshold as 0.5.

We evaluate the performance of the predictive models by performing a five-fold cross-validation and measuring the area under the receiver operating curve (AUC) generated by plotting sensitivity vs $(1 - \text{specificity})$. We use cross-validation for detecting and preventing possible overfitting or selection bias. We randomly divided the dataset into five subsamples, retained a single subsample as the validation data for testing the model, and the remaining four samples used as training data. We repeated the process five times (the folds), with each of the subsamples used exactly once as the validation data. Performance of the model in each fold was measured and then results from all five folds were averaged to produce a single estimation for the model's performance.

AUC measures the overall performance of the recall with respect to different false positive rate. Models with higher AUC will demonstrate a more powerful screening capability in assisting the physicians. In order to further evaluate the machine learning models for a clinical setting, we assess the AUC along with operating points corresponding to high-sensitivity (true positive rate) and high-specificity (true negative rate) of the algorithm with respect to the reference standard [174, 175, 176, 177, 178]. Targeted operating points are used for different clinical purposes, for instance high-sensitivity is targeted for ruling out the disease, whereas high-specificity is used for ruling in the disease [179]. In this study, in order to evaluate the performance under consistent conditions, the operating points correspond to fixed sensitivity and specificity at 0.80 and 0.85 [174, 176, 177]. In practice, high-sensitivity (or recall rate of positive cases) plays a more important role in screening the patients. In essence, a highly sensitive test indicates that the model can

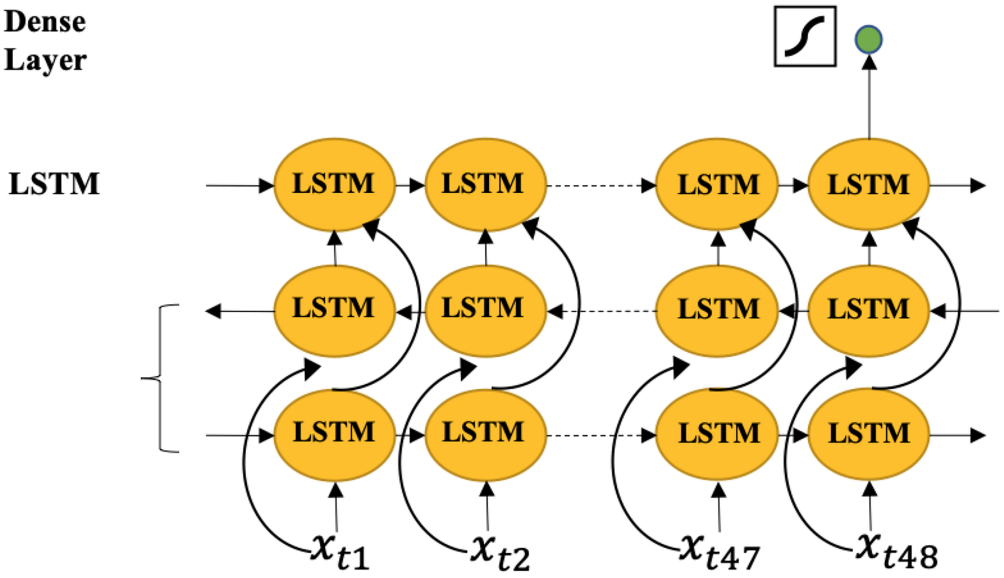

Figure 5.4: LSTM model. A bidirectional LSTM combined with an additional LSTM layer, followed by a dense decision layer with one output neuron activated by a sigmoid function. Overall, we have 16 hidden units in our LSTM layer.

correctly identify patients with a high risk of readmission in a critical department such as ICU.

5.3 RESULTS

In this section, we illustrate the experiments we conducted to evaluate the performance of the predictive models. We evaluate the conventional models (logistic regression, random forest, Naive Bayes, and SVM), as well as, the deep learning based CNN and temporal LSTM models. We compare and obtain the optimal ICU readmission prediction solution.

5.3.1 Baseline models

We first evaluate logistic regression models with both *L1* and *L2* regularization penalty. Results are shown in Table 5.3 part (a) under "Baseline-Regression". We first observe that using logistic regression with L2 regularization on the advanced statistical features (A_STAT) can slightly improve the AUC performance compared to the basic statistical features (B_STAT), 0.770 (95% CI, 0.758-0.782) to an AUC of 0.771 (95% CI, 0.759-0.783). However, we do not observe any AUC improvement for the logistic regression with L1 regularization by having more advanced statistical features (stays at AUC of 0.775 (95% CI, 0.765-0.786)). In addition to advanced statistical features, the demographic features can also slightly improve the performance from an AUC of 0.771 (95% CI, 0.759-0.783) to an AUC of 0.773 (95% CI, 0.762-0.787) using the logistic regression with L2 regularization.

Table 5.3: Performance comparison of various machine learning models on different sets of features. Acc: Accuracy. Pre: Precision. Re: Recall. A.R: AUC under ROC. A.P: AUC under PRC. L48: Last 48 hours. F48: First 48 hours. CE: Chart Events. D: Demographic features. C.I.: 95% confidence interval. B_STAT (basic statistical features): slope and intercept. A_STAT (advanced statistical features): B_STAT plus continues and categorical features including quadratic term, standard deviation, mean absolute deviation, R^2, Majority value, value change frequency.

Begin of Table 5.3			
Model	**Features**	**Re-1 (95% CI)**	**A.R (95% CI)**
(a) Baseline - Regression			
LR-L2	L48-h B_STAT + ICD9	0.67 (0.647 - 0.694)	0.77 (0.758 - 0.782)
LR-L2	L48-h A_STAT + ICD9	0.67 (0.648 - 0.692)	0.771 (0.759 - 0.783)
LR-L2	L48-h A_STAT + ICD9 + D	0.676 (0.650 - 0.703)	0.773 (0.762 - 0.787)
LR-L1	L48-h B_STAT + ICD9	0.669 (0.647 - 0.691)	0.775 (0.764 - 0.786)
LR-L1	L48-h A_STAT + ICD9	0.669 (0.656 - 0.681)	0.775 (0.765 - 0.786)
LR-L1	L48-h A_STAT + ICD9 + D	**0.68 (0.662 - 0.697)**	**0.777 (0.765 - 0.789)**
(b) Baseline - Conventional Machine Learning			
NB	L48-h B_STAT + ICD9 + D	0.453 (0.434 - 0.472)	0.709 (0.702 - 0.716)
NB	L48-h A_STAT + ICD9 + D	0.509 (0.479 - 0.540)	0.706 (0.698 - 0.713)
RF	L48-h B_STAT + ICD9 + D	0.563 (0.548 - 0.578)	0.714 (0.703 - 0.725)
RF	L48-h A_STAT + ICD9 + D	0.565 (0.550 - 0.580)	0.712 (0.693 - 0.730)
SVM	L48-h B_STAT + ICD9 + D	0.701 (0.686 - 0.715)	0.775 (0.765 - 0.785)
SVM	L48-h A_STAT + ICD9 + D	**0.703 (0.685 - 0.720)**	**0.779 (0.768 - 0.789)**

Table 5.3 (cont.)			
Model	**Features**	**Re-1 (95% CI)**	**A.R (95% CI)**
(c) Feature Selection			
LSTM	F48-h CE + ICD9	0.731 (0.723 - 0.740)	0.777 (0.769 - 0.785)
LSTM	L48-h CE + ICD9	0.717 (0.692 - 0.742)	0.784 (0.772 - 0.795)
LSTM	L48-h CE	0.593 (0.537 - 0.649)	0.704 (0.697 - 0.710)
LSTM	L48-h CE + ICD9 + D	**0.733 (0.698 - 0.768)**	**0.787 (0.771 - 0.802)**
(d) Model Selection			
CNN	L48-h CE + ICD9	0.665 (0.586 - 0.745)	0.78 (0.774 - 0.786)
CNN	L48-h CE + ICD9 + D	0.735 (0.676 - 0.794)	0.784 (0.773 - 0.794)
CNN+LSTM	L48-h CE + ICD9	0.739 (0.670 - 0.807)	0.785 (0.775 - 0.795)
CNN+LSTM	L48-h CE + ICD9 + D	0.71 (0.648 - 0.771)	0.787 (0.775 - 0.799)
LSTM+CNN	L48-h CE + ICD9	0.729 (0.647 - 0.811)	0.786 (0.776 - 0.796)
LSTM+CNN	L48-h CE + ICD9 + D	**0.742 (0.718 - 0.766)**	**0.791 (0.782 - 0.800)**
End of Table 5.3			

Overall, we see that the prediction accuracy can be slightly improved by adding more complex statistical features as well as demographic ones. The best performing logistic regression model is with L1 regularization on A_STAT combined with the demographic features, AUC of 0.777 (95% CI, 0.765-0.789) and sensitivity of 0.680 (95% CI, 0.662-0.697). Furthermore, we trained three conventional machine learning models as our baseline, including Naive Bayes, Random Forest, and SVM on both B_STAT and A_STAT features. The results are shown in Table 5.3, part (b), "Baseline-Conventional Machine Learning". SVM outperforms other traditional methods by reaching an AUC of 0.779 (95% CI, 0.768-0.789) with A_STAT, which is a negligible increase from an AUC of 0.775 (95% CI, 0.765-0.785) with B_STAT.

5.3.2 CNN and LSTM models

We first conduct a feature ablation study to evaluate the effect of various feature selections on the system's performance. Then, we attempt multiple model structures including bidirectional LSTM, CNN, and the combinations of both.

Feature selection: We select the Bidirectional LSTM as our base model and deploy different combinations of feature inputs. As shown in Table 5.3 part (c), our results demonstrate that the last-48h features perform relatively better than the first-48h data in terms of positive recall rate and AUC. In addition, ICD-9 embedding is necessary for predicting the readmission rate. We also observe that the demographic features greatly benefit the performance. Overall, the full set of features including Last-48h chart events and their identifiers, ICD-9 embeddings, and demographic information perform the best among all the combinations.

Model selection: We attempted multiple model structures including bidirectional LSTM, CNN, and the combinations of both. Figure 5.5 shows our strategy for combining the bidirectional LSTM and CNN models. We use the 1D multi-filter CNN model introduced in the previous section. As for the CNN+LSTM model, the CNN follows a multi-filter convolution computation with zero padding to maintain the timestamp consistency for different groups of feature maps. The following LSTM only outputs the hidden units of the last time stamp. However, for the LSTM+CNN model, CNN computes the feature maps without zero padding after receiving the output hidden unit sequence from LSTM. As shown in Table 5.3 part (d), our experimental results reveal that LSTM followed by a CNN, utilizing all the feature sets, obtains a higher positive recall rate and overall prediction performance. The proposed model outperforms the conventional machine learning approaches trained on both basic and advanced statistical features. The ROC curve for some of the selected high performing machine learning models are shown in Figure 5.6.

To further demonstrate the ability of deep learning model in the readmission prediction, we look at the operating points corresponding to high-sensitivity (true positive rate) and high-specificity (true negative rate) of the algorithm. Table 5.4 summarizes the performance of the algorithms. Using the operating cut point with high specificity of 0.85 and 0.8, we observe that LSTM+CNN results in the highest sensitivities of 0.548 (95% CI, 0.522-0.575) and 0.619 (95% CI, 0.597-0.642) respectively, a significant improvement from the best baseline. Evidently, even the basic LSTM model outperforms the best baseline, regression with L1 regularization, by improving the sensitivities from 0.525 (95% CI, 0.505-0.546) to 0.540 (95% CI, 0.503-0.577) and 0.596 (95% CI, 0.575-0.618) to 0.611 (95% CI, 0.573-0.649) respectively.

We then evaluated a second operating point for the algorithm, with a high-sensitivity, reflecting an output that would be used for a screening tool. Using this operating point, LSTM+CNN had

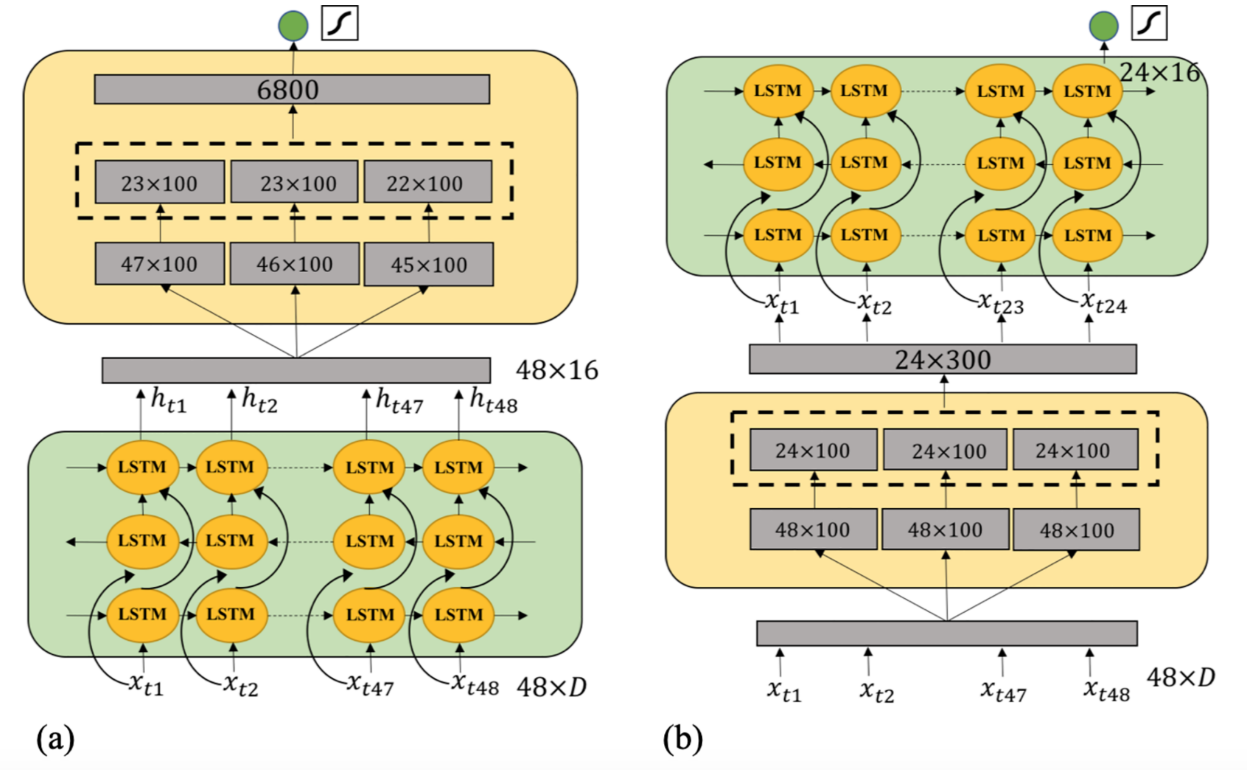

Figure 5.5: Combination of LSTM and CNN models. (a) CNN+LSTM model, the CNN follows a multi-filter convolution computation with zero padding to maintain the timestamp consistency for different groups of feature maps. The following LSTM only outputs the hidden units of the last time stamp. (b) LSTM+CNN model, CNN computes the feature maps without zero padding after receiving the output hidden unit sequence from LSTM.

sensitivities of 0.85 and 0.8 and the highest specificities of 0.537 (95% CI, 0.515-0.559) and 0.618 (95% CI, 0.593-0.643), again an improvement from conventional machine learning models.

5.4 DISCUSSION

In this section, we dive deeper into our machine learning model in an effort to further interpret the results, its capabilities, and limitations. We perform ablation study to investigate the most important factors that the deep learning model has learned in order to predict the ICU readmission. Then, we review the clinical literature for additional verification and a better clinical understanding of the deep learning model. Finally, we examine the advantages and strength of the proposed model over traditional machine learning models. We look at the characteristics and statistics for the true positive sets of each model.

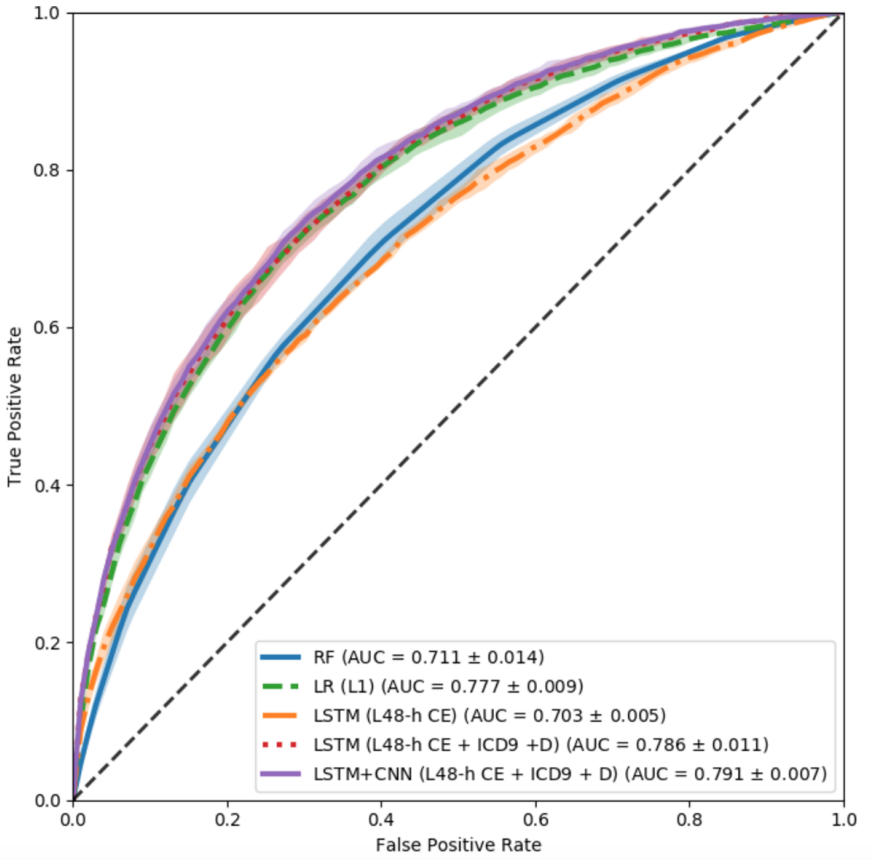

Figure 5.6: ROC curve of selected high performing machine learning models. The color bar is the error bar of the ROC curve with five-fold cross-validation. LSTM-CNN model performs relatively better than other ones. CE: chart events. D: Demographic features.

5.4.1 Model interpretation: Feature ablation test

We conducted the feature ablation test on the chart events to better understand the underlying logic of our proposed model. We selected all the positive cases from the testing partition. Then we obtained all the true positive samples through running the LSTM+CNN model utilizing all the features. These true positive cases are the ones recalled correctly by our proposed model. For each case, we iterated over all the chart events, each time, changing only one event to its normal value in the humans. We recorded the number of cases that were falsely predicted due to the change. Then we ranked all the chart events according to the change numbers. Figure 5.7 shows the results of feature ablation test based on the changing ratio of the prediction results after we replace the original feature with its normal value. We see that Glucose is the most important factor learned by the deep learning model for the readmission prediction task, while Capillary Refill Rate, Fraction inspired Oxygen, and Systolic Blood Pressure do not have significant influence on the prediction results. However, the performance change of the predictive model is not dramatic. We believe this may be due to possible biological and clinical correlation among different factors. This can be

Table 5.4: Performance comparison of machine learning models at high-sensitivity and high-specificity operating points.

Model	**Re-1 (95% CI) (Re-0 fixed near 0.85)**	**Re-1 (95% CI) (Re-0 fixed near 0.8)**	**Re-0 (95% CI) (Re-1 fixed near 0.85)**	**Re-0 (95% CI) (Re-1 fixed near 0.8)**
Baseline - Regression				
LR-L2	0.507 (0.491 - 0.522)	0.59 (0.570 - 0.611)	0.516 (0.484 - 0.549)	0.596 (0.563 - 0.629)
LR-L1	**0.525 (0.505 - 0.546)**	**0.596 (0.575 - 0.618)**	**0.518 (0.476 - 0.561)**	**0.599 (0.573 - 0.626)**
Baseline - Conventional Machine Learning				
NB	0.358 (0.333 - 0.383)	0.468 (0.447 - 0.489)	0.247 (0.238 - 0.256)	0.33 (0.318 - 0.342)
RF	0.402 (0.364 - 0.439)	0.475 (0.445 - 0.504)	0.415 (0.381 - 0.449)	0.487 (0.457 - 0.516)
SVM	**0.519 (0.498 - 0.539)**	**0.596 (0.584 - 0.608)**	**0.532 (0.498 - 0.565)**	**0.608 (0.588 - 0.628)**
Deep Learning				
LSTM	0.54 (0.503 - 0.577)	0.611 (0.573 - 0.649)	0.532 (0.503 - 0.561)	0.608 (0.590 - 0.626)
CNN	0.531 (0.513 - 0.549)	0.604 (0.579 - 0.630)	0.527 (0.493 - 0.561)	0.607 (0.573 - 0.641)
CNN+LSTM	0.543 (0.510 - 0.576)	0.617 (0.590 - 0.644)	0.535 (0.515 - 0.556)	0.611 (0.591 - 0.632)
LSTM+CNN	**0.548 (0.522 - 0.575)**	**0.619 (0.597 - 0.642)**	**0.537 (0.515 - 0.559)**	**0.618 (0.593 - 0.643)**

further evaluated by the back-propagation approach in future work.

5.4.2 Model interpretation: Features in line with the clinical literature

Furthermore, we review the clinical literature for additional verification and a better understanding of the deep learning model system. The results of the feature ablation test from the previous section point out that abnormal Glucose, Heart Rate, Body Temperature, Glasgow Coma Scale, and Oxygen Saturation are the top five important features in predicting unplanned readmission in the ICU. Interestingly, the underlying deep learning logic and its findings are in line with the existing clinical literature. Prior research has found that the presence of comorbidities, such as diabetes, heart failure, renal failure, and pneumonia, are the main risk factors resulting in unplanned readmissions [180, 181]. These disorders are shown to have strong correlations with abnormal features identified by our model [157]. Moreover, several studies have worked on the readmission problem by only focusing on the aforementioned conditions.

For instance, many researchers have focused on hospitalization and unplanned readmissions by looking at the abnormal Glucose status. Berry et al discovered the significant positive relationship between levels of admission blood glucose and risk of readmission for patients with heart failure [182]. Evans et al identified that the glucose level on admission performs as a prognostic predictive factor for early readmission rates, even for those with diabetes [183]. Dungan has demonstrated that higher time-weighted mean glucose is associated with the increase of congestive heart fail-

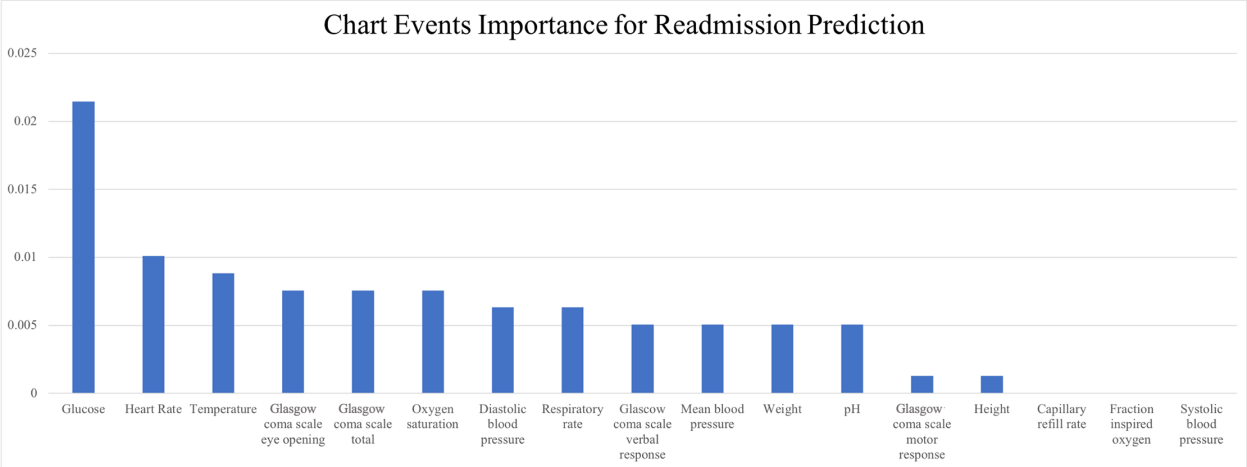

Figure 5.7: The results of feature ablation test. The importance of chart events for predicting the ICU readmission. The y-axis shows the changing ratio of the prediction results after we replace the original feature with its normal value.

ure (CHF) readmission [184]. Emons et al focused on hypoglycemia-related readmission issue and expose the linear relationship between blood glucose level closest to discharge and the risk of hypoglycemic readmission [185].

Heart failure is another main risk factor resulting in early readmission [186]. Heart failure indicates that the cardiac muscle cannot pump the blood properly. This behavior is strongly reflected through abnormal heart rate [187, 188]. Keenan et al developed a hierarchical logistic regression model to predict readmission for those patients hospitalized with heart failure issues [187]. Hammill et al utilize heart rate record during hospitalization as one of the main features to predict 30-day outcomes after heart failure hospitalization [188].

In addition, patients with renal failure are suggested to be among the highest risk patients with 30-day readmission [189]. Previous studies have shown that body temperature is a vital determinant of ischemic renal injury [190]. Moreover, Sood et al found that body temperature and Glasgow coma scale are two significant features to predict early ICU readmission for patients with end-stage renal disease (ESRD) [191].

Last but not least, a study has revealed that around 140,000 hospital readmissions per year are owing to pneumonia [192]. Halm et al apply a regression to examine the relationship between patients' instabilities and the risk of early readmission. They proposed a list of unstable factors leading to higher risk of 30-day hospital readmission, including (temperature >37.8°C, heart rate >100 bpm, respiratory rate >24/min, systolic blood pressure <90 mmHg, oxygen saturation <90%, inability to maintain oral intake, and abnormal mental status) [193].

In summary, the underlying logic of our deep learning model, as well as the most important features identified by the model, are in line with the existing clinical literature.

5.4.3 Strengths of the model

To better understand the advantages and strength of the LSTM-based model over the traditional models, we investigate the positive patients correctly predicted by the LSTM+CNN but misclassified by the logistic regression with L1 regularization. Overall, there are 441 positive patients, across all the testing partition folds, who are correctly predicted only by the LSTM+CNN model and not the logistic regression. We refer to these 441 patients as LSTM-C set. Meanwhile, 3,068 cases are correctly predicted by both the LSTM+CNN and Logistic Regression with L1 regularization. We refer to these 3,068 cases as LSTM-LR-C set.

LSTM-based models are found to provide a robust prediction for time series with notable fluctuations in the data [194]. We verify this phenomenon by measuring the degree of value oscillation for LSTM-C and LSTM-LR-C, and also looking at individual cases. We introduce D_{nm}, measuring the degree of oscillation for record n of chart event m. Given a numerical chart event sequence $E_{nm} = \{x_t\}$, where $t \in [1, 48]$, then D_{nm} can be computed by,

$$D_{nm} = \frac{1}{T-1} \sum_{t=2}^{T} |x_t - x_{t-1}| \tag{5.3}$$

where T is equal to the length of a record, normally 48, if there is no missing data.

Using D_{nm} as a measure for the degree of oscillation, we compute the highest oscillation for each stay across all the 12 numerical chart events and compare their statistical distributions in LSTM-C and LSTM-LR-C. We first estimate P_m, the cumulative density function (CDF) of each chart event on the whole positive set. Then we remapped each D_{nm} to the probability p_{nm}, and computed the maximum probability w_n for each record n by,

$$\begin{aligned} p_{nm} &= P_m(D_{nm}) \\ w_n &= \max_m p_{nm} \end{aligned} \tag{5.4}$$

where w_n represents the highest oscillation among all the chart events for this record.

Finally, for both LSTM-C and LSTM-LR-C sets, we plotted the CDFs of the estimated histograms of w_n in Figure 5.8. We can see that there are more patient records in the LSTM-C which have at least one chart event with high oscillation sequence. Essentially, compared to Logistic Regression, our LSTM+CNN model is capable of capturing high volatile time series behavior, a

common pattern in high-risk ICU patients.

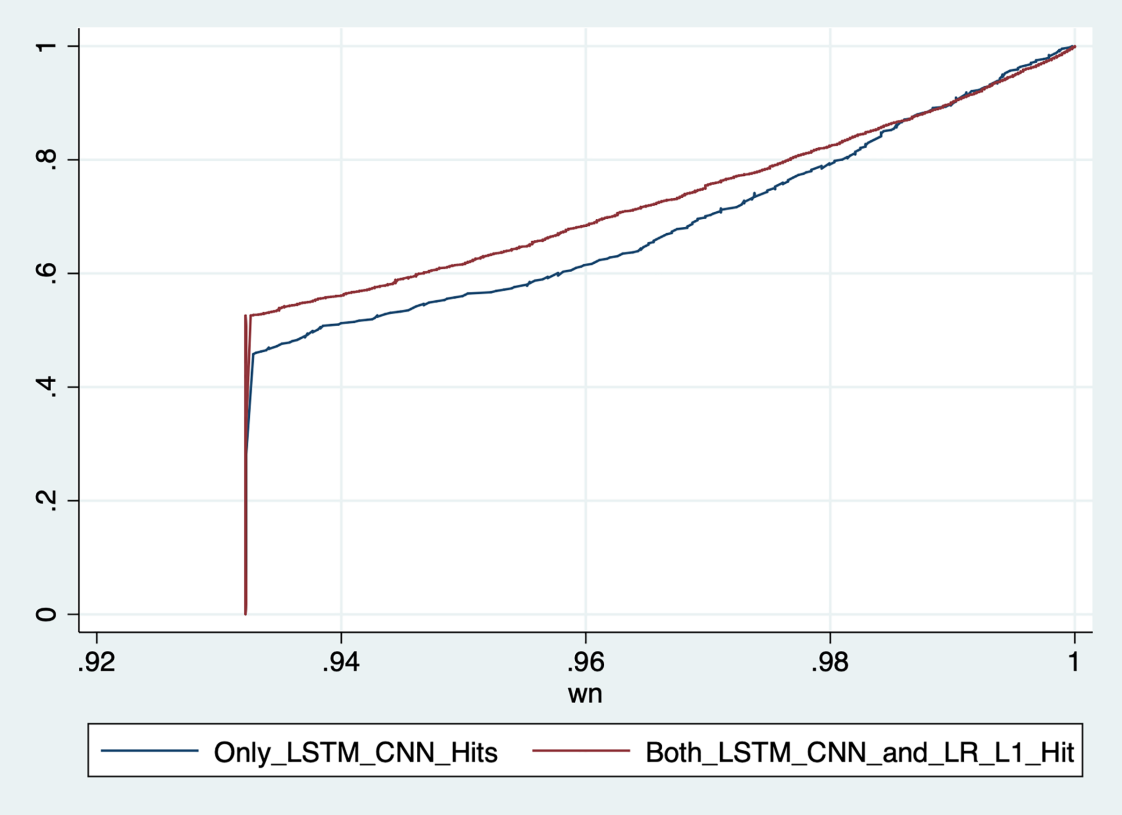

Figure 5.8: Cumulative density function curve of LSTM-LR-C (red line) and LSTM-C (blue line). Figure shows that there are more patient records in the LSTM-C which have at least one chart event with high oscillation sequence. Essentially, compared to Logistic Regression, our LSTM+CNN model is capable of capturing high volatile time series behavior, a common pattern in high-risk ICU patients.

To further study the strength of our LSTM+CNN solution, we look at individual cases. For each chart event, we selected the patients with the highest D_{nm} in the LSTM-C and plotted the sequence values of their stay. Figure 5.9 illustrates two of these patients. Both patients have high volatile chart events. However, in both cases the abnormal sequence has oscillated around the normal value of the chart event, which in return a linear model would regress it to a normal value with a negligible slope. Effectively, a linear model would lose a very important factor in predicting the readmission: repeated illness and unstable status.

We further investigate the strength and weaknesses of the LSTM-based model by looking at the oscillation issue among all the chart events. We investigate the differences between positive patients who are predicted correctly only by the logistic regression with L1 regularization and those

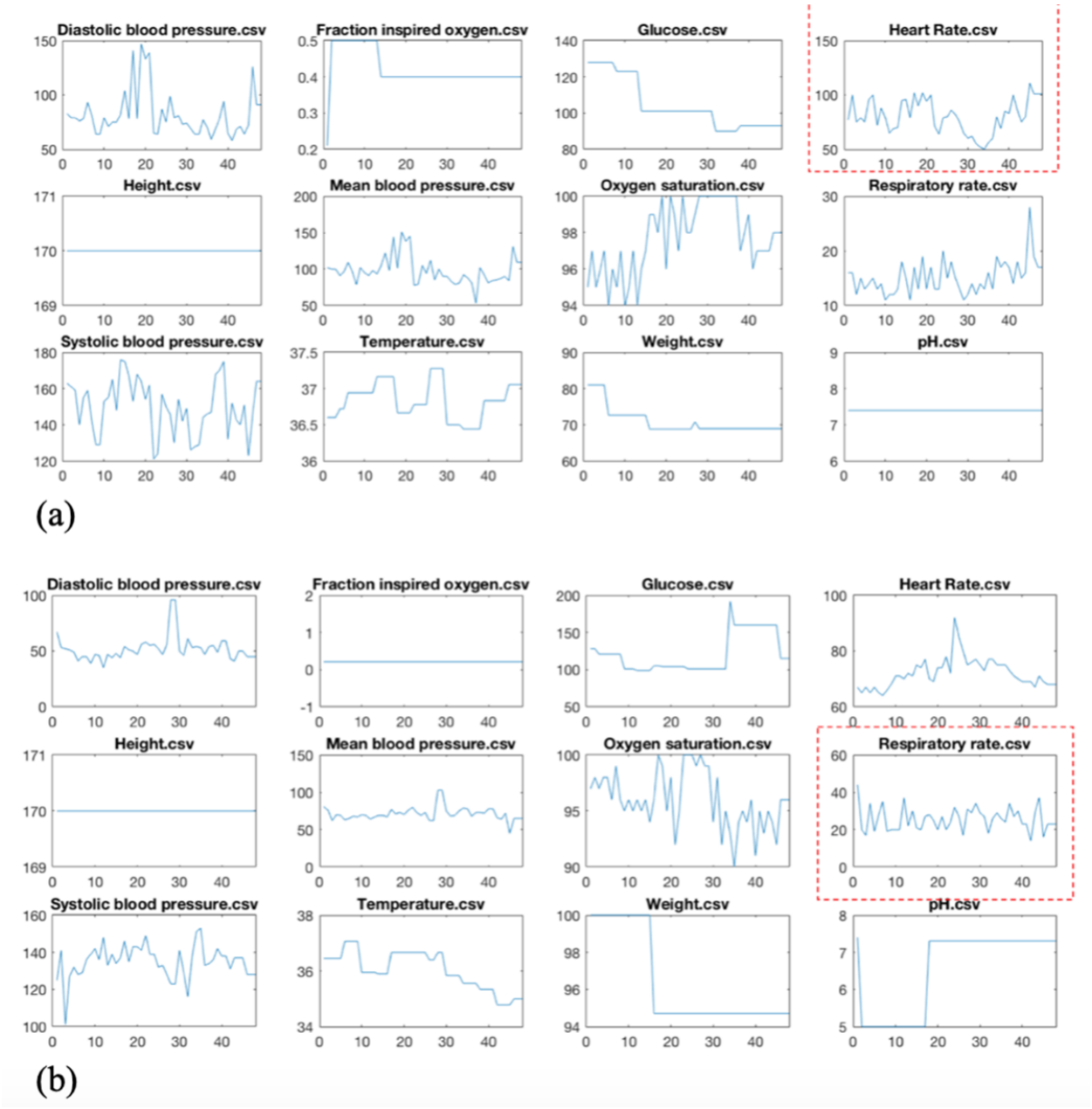

Figure 5.9: (a) A selected ICU stay with the highest heart rate event oscillation, and (b) another case with the highest oscillation of respiration rate. These two patients are predicted correctly by the LSTM-CNN model, but wrongly by the traditional models. In both cases, the abnormal sequence has oscillated around the normal value of the chart event, which in return a linear model would regress it to a normal value with a negligible slope. Effectively, our LSTM-CNN is capable of capturing such high volatile behavior, a common pattern among high-risk ICU patients with unstable status.

Table 5.5: KolmogorovSmirnov (K-S) test for the distribution of fluctuation between LSTM-C and LR-C for each chart event.

Two-sample KolmogorovSmirnov (K-S) test		
	D	**P value**
Glucose	0.1519	0.012
Heart rate	0.0635	0.766
Temperature	0.0839	0.42
Oxygen saturation	0.1678	0.004
Diastolic blood pressure	0.0794	0.491
Respiratory rate	0.102	0.201
Mean blood pressure	0.068	0.687
Weight	0.0476	0.964
pH	0.0635	0.766
Height	0.0181	1
Fraction inspired oxygen	0.0499	0.947
Systolic blood pressure	0.0635	0.766

who are predicted correctly only by the LSTM+CNN. As mentioned earlier, there are 441 positive patients who are predicted correctly only by the LSTM+CNN model, denoted as the LSTM-C set. On the other hand, there are 147 cases that are predicted correctly only by the logistic regression with L1 regularization, we denote this set by LR-C.

Our goal is to identify the differentiating factors between the LR-C set and the LSTM-C. We analyze the fluctuation distribution for each chart event in both sets. We use the Eq 5.3 to calculate the D_{nm}, measuring the degree of oscillation for chart event m of patient n. We then estimated the cumulative density function (CDF) of each chart event in each set. For each chart event, we conduct KolmogorovSmirnov test (K-S test) on factor D_{nm} to compare the distributions of this factor between the two sets. The results are shown in Table 5.5.

Results reveal that patients in the LR-C tend to have a higher probability of achieving lower scores of factors Dnm on "Glucose" than patients in the LSTM-C (maximal absolute difference between the distribution functions (D) = 0.1519, p-value = 0.012). In addition, we also observe that patients in the LR-C set tend to have a higher probability of achieving lower scores of factors D_{nm} on "Oxygen Saturation" than patients in LSTM-C (D = 0.1678, p-value = 0.004). The CDF of "Glucose" and "Oxygen Saturation" are shown in Figure 5.10 part (a) and (b). We further use the Probability density function (PDF) plots of both features to show this phenomenon in Figure 5.10 part(c) and (d). The results in this section further enhance the suggestion that deep learning has advantages over logistic regression in predicting datasets with large fluctuation of time series features, "Glucose" and "Oxygen Saturation" in this case.

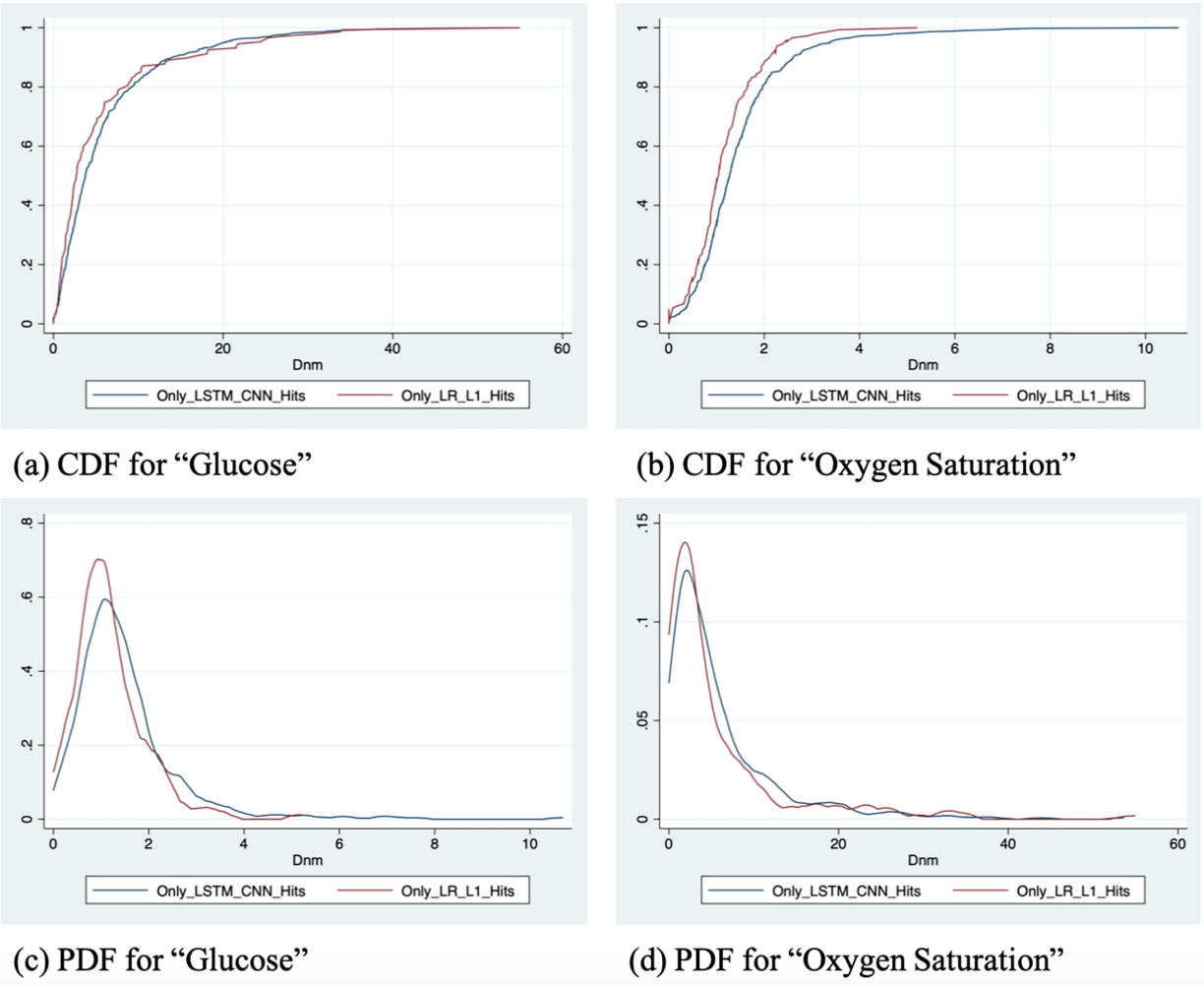

Figure 5.10: Cumulative density function (CDF) plots and probability density function (PDF) plots.

5.4.4 Comparison with the baseline models

In addition to comparing LSTM-based model and logistic regression with L1 regularization, we further compare LSTM-based model with all the six baseline regression models, including: (i) L1 logistic regression with B-STAT, (ii) L1 logistic regression with A_STAT, (iii) L1 logistic regression with A_STAT and Demographic features, (iv) L2 logistic regression with B-STAT, (v) L2 logistic regression with A_STAT, and (vi) L1 logistic regression with A_STAT and Demographic features.

We define LSTM-C-all as the set of positive ICU readmission cases which can only be identified by LSTM+CNN model and not any of the six baseline regression models as mentioned above. Overall, 201 cases are contained in LSTM-C-all.

Table 5.6: Summary of the number of cases correctly predicted by the corresponding baseline model as well as the LSTM+CNN.

Name of the set	**Number of cases**
LSTM-C-all	201
LSTM-LR-L1-B_STAT	3,033
LSTM-LR-L1-A_STAT	3,022
LSTM-LR-L1-A_STAT-D	3,068
LSTM-LR-L2-B_STAT	3,044
LSTM-LR-L2-A_STAT	3,079
LSTM-LR-L2-A_STAT-D	3,084

Furthermore, we define the following sets: (i) LSTM-LR-L1-B_STAT set, (ii) LSTM-LR-L1-A_STAT set, (iii) LSTM-LR-L1-A_STAT-D set, (iv) LSTM-LR-L2-B_STAT set, (v) LSTM-LR-L2-A_STAT set, (vi) LSTM-LR-L2-A_STAT-D set as the sets that are correctly predicted by both the LSTM+CNN and respective baseline logistic regression models. Summary of the number of cases contained in each of these sets is shown in Table 5.6.

We follow the same logic described in the previous section, to capture the maximum probability w_n of record n for each set mentioned above. Then, we plot the CDF of w_n for each set. Results are shown in Figure 5.11. Figure shows that the CDF representing oscillation of the LSTM-C-all set is still the lowest one. The observation is consistent with the previous observation in the section "Strengths of the model": there are more patient records in the LSTM-C-all which have at least one chart event with high oscillation sequence. The result enhances our argument that compared to baseline logistic regression models, our LSTM+CNN model is capable of capturing high volatile time series behavior, a common pattern in high-risk ICU patients.

The rest of the CDF lines represent oscillation of the six sets predicted correctly by both the LSTM+CNN and various logistic regressions models. We observe that the six CDFs are almost identical. Based on this observation, we conclude that even though using A_STAT (mode advanced statistical features) can slightly enhance the performance of baseline logistic regression models (as shown in Table 5.3), it can hardly improve the ability of logistic regressions to capture the critical oscillations in ICU patients. The results further enhance the advantage of using LSTM based model to identify patients with a high risk of readmission in a critical department such as ICU.

5.5 SUMMARY

In this study, we addressed the unplanned ICU readmission prediction by utilizing chart events, demographics and ICD-9 embeddings features. Among the data that we used, chart event features are significantly sensitive to time series, and cannot be properly captured by conventional machine

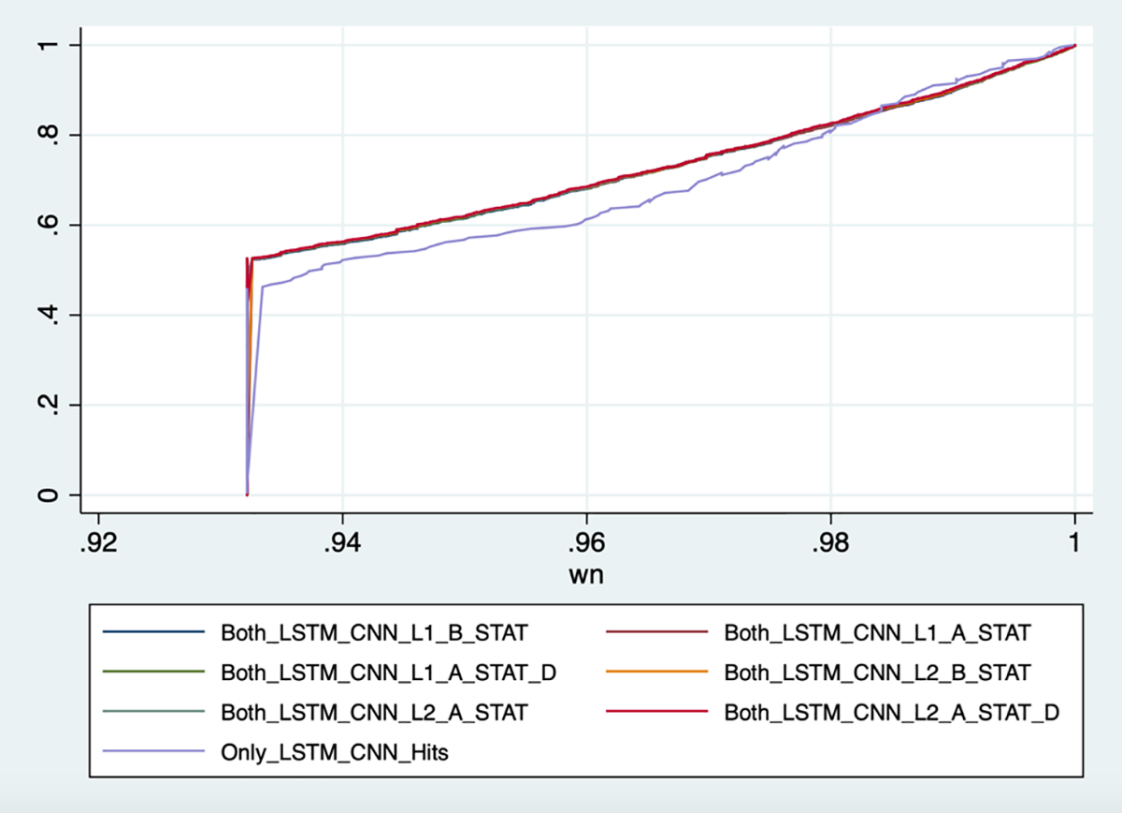

Figure 5.11: CDF plots of w_n for sets in Table 5.6.

learning models (e.g., logistic regression). After in-depth analysis of ICU data, it turned out that sudden fluctuations in chart event features (e.g. glucose and heart rate) are strong signals of patients volatile status. We propose a LSTM-CNN based model, which can properly incorporate time series data without information lost.

Our machine learning solution for prediction ICU readmission offers higher accuracy and sensitivity compared to existing solution. In addition, since the model can have multiple operating points, its sensitivity and specificity can be tuned to match requirements for specific clinical settings, such as high sensitivity for critical care. In this study, AUC of 0.791 and sensitivity of 0.742 were achieved. Moreover, we illustrated the importance of each input features and their combinations in the predictive model This fast and interpretable solution holds the potential for substantial clinical impact by augmenting clinical decision-making for ICU specialists. Further research is necessary to evaluate performance in a real-world, clinical setting, in order to validate this technique across varying critical care practices.

CHAPTER 6: CONCLUSIONS

6.1 SUMMARY OF CONTRIBUTIONS

In summary, this work used machine learning techniques to enable the construction of the human aging model. Descriptive, predictive, prescriptive machine learning methods contributed to the human aging model and permitted health-related decisions for clinicians and individuals. We have taken a comprehensive and multi-disciplinary approach to the human aging model described by three principal components: (i) healthy aging, (ii) longitudinal and multi-modal data, (iii) machine learning and cloud computing. We focus on brain aging and neurological disorders. This subset of aging-related disorders is representative of a large portion of aging trajectories. Our data-driven strategy starts with the collection and integration of longitudinal studies with highly dimensional multi-modal datasets. With longitudinal and multi-modal data in hand, we utilize machine learning techniques including supervised, unsupervised, and semi-supervised learning to analyze the data and develop models. Analyzing health data and particularly imaging and genomics data, pose severe computational challenges. To address the storage and computational challenges posed by big health data, we design and use cloud computing-based analytical systems that have transitioned from shared, centralized architectures to distributed, decentralized architectures. We propose several models for aging and neurological disorders as one of the most prominent health disorders in our aging population. These models address Parkinson's disease, Alzheimer's disease, Amyotrophic lateral sclerosis (ALS), and Intensive Care Unit (ICU). Our solutions impact the whole spectrum of aging-related healthcare from patients and caregivers to physicians and clinicians to providers and insurers. This work complements and supports many other researchers' efforts in modeling aging-related disorders.

In Chapters 2 and 3, we studied Parkinson's and Alzheimer's disorders, respectively. In both chapters, we integrated unlabeled, multi-modal, and longitudinal data. The longitudinal data had a long-term nature, and we were interested in capturing the overall pattern of the individual's trajectories. Vectorization and NMF methods were the most successful approach for extracting long-term trajectories. Using comprehensive multi-modal data helped us develop an embedded space. This space was crucial for understanding the trajectories and dimensions in which the individuals traverse. Having this easily interpretable space, we were able to use GMM unsupervised learning approach to identify new subtypes of the disorder based on disease progression. We also provided an in-depth analysis of these subtypes. Furthermore, we developed predictive models for early diagnosis, prognosis, and clinical trial stratification. Significant effort was also put into the interpretation of the predictive models. Predicting progression rates would lead to better patient-specific attention by recognizing at an early stage the patients with a swift rate

of progression. The proposed disease progression and trajectory prediction algorithms can help doctors and practitioners develop a methodical and organized course for clinical tests, which can be much more concise and effective in detection. These adaptations and modifications in clinics may help to diminish treatment and therapy costs for aging disorders. Further, the capability to anticipate the trajectory of impending Alzheimer's and Parkinson's progression at the early stages of the disease is an advancement towards uncovering novel treatments for disease modification. The proposed analysis provides insights to inhibit or decelerate the progression of aging-related symptoms and subsequent deterioration in the characteristics of life that are accompanied by the disease.

In Chapter 4, we introduced our work on Amyotrophic lateral sclerosis (ALS) disorder. Similar to Alzheimer's and Parkinson's chapters, we used an unsupervised learning approach to identify new subtypes. However, unlike Parkinson's and Alzheimer's, which have a long multi-year course, we do not predict progression. Since ALS is a rapidly progressive disorder, we focus on subtype identification and survival analysis. To enhance our ability to resolve the different types of spinal-onset ALS, we used semi-supervised learning. We incorporated the physician's assessment of patients for the predictive task. With the trained model, we dissected the machine learning model to realize the underlying ALS structure which machine has learned. This representation is a close proxy to what the model learned about data for it to classify patients as observed by the physician a year into the disease – resulting in a more fine-tuned representation of ALS subtypes compared to the fully unsupervised. Notably, the machine learning combined with the physician's observations arrived at a more fine-tuned conclusion, highlighting the importance of augmentation (and not a replacement) of the machine with clinicians. In order to ensure the generalizability and validity of the results, we replicate the ALS subtype identification in the independent replication cohort.

In Chapter 5, we presented our work on readmission prediction in the Intensive Care Unit (ICU). Unlike previous chapters where the health issue has a multi-year course, in the ICU, data is short-term. Short-term data such as sensory and monitoring device data has a high frequency of data collection, but with a time window of hours to days. Due to the data collection frequency, different predictive machine learning methods are used to encapsulate the time-sensitivity of small fluctuations in the patient's status. After an in-depth analysis of ICU data, it turned out that sudden fluctuations in chart event features (e.g., glucose and heart rate) are strong signals of patients' volatile status. Comparing various methods, we show that our LSTM-based solution can better capture high volatility and unstable status in ICU patients, an important factor in ICU readmission. This work highlights the ability of machine learning models to improve our ICU decision-making accuracy and is a real-world example of precision medicine in hospitals. Due to human limitations, many behavioral changes and sudden fluctuations are most often missed. Moreover, we illustrated the importance of each input features and their combinations in the predictive model. This fast and interpretable solution holds the potential for substantial clinical impact by augmenting clinical

decision-making for ICU specialists.

The models and solutions developed in this dissertation are designed for different scenarios, but share the following designing principles:

i *Lack of data for supervised learning.* In order to develop accurate predictive models based on supervised learning, we need large and reliable data. In healthcare, such data is mostly not available; labels come from physicians who themselves have a high misdiagnosis rate. To overcome this challenge, this work heavily relies on labeling data using *unsupervised* and *semi-supervised learning* techniques.

ii *Utilization of short-term and long-term longitudinal data.* In order to develop a timely model of aging, we need to incorporate longitudinal studies with time-series data. In healthcare, we have both *short-term* and *long-term* longitudinal data. Short-term data such as sensory and monitoring device data has a high frequency of data collection, but with a time window of hours to days. On the other hand, long-term data such as clinical assessments have a low frequency of data collection with a time window of six months to two years. In this work, we utilize both types of longitudinal data.

iii *Integration of multi-modal data.* Many machine learning techniques focus on one type of data, e.g., only imaging or audio. However, in healthcare, we look at the human body from different modalities, and we hope by integrating all of them, we can have a better understanding of the issue. Utilizing highly dimensional multi-modal datasets, including clinical, biological, genetic, and imaging data have been part of this work.

iv *Interpretability.* Using unsupervised machine learning techniques, we have developed *embedding spaces* that have guided us in labeling the subjects. Part of this labeling relies on our success in interpreting the embedding spaces. In this work, we have put effort into dissecting the machine learning "black box" to understand the results better and guide the development of models.

We also address the following clinical challenges:

i *Replication of results with other datasets.* Validating findings and replication is an underpinning of research. Generalizing, the same methods and protocols should be used on a different group of people, or a different setting, and come up with similar results. In healthcare, lack of data has made replication more sparse. However, in this work, when possible, we show that results are valid in external datasets.

ii *Developing usable models in both clinical and research settings.* Models with higher accuracy demonstrate a more powerful screening capability in assisting the physicians. However, accuracy is not enough, and we need to further evaluate the machine learning models for use in a

clinical setting. We need to assess the accuracy along with operating points corresponding to sensitivity (also called the true positive rate, the recall, or probability of detection) and specificity (true negative rate) of the algorithm with respect to the reference standards [61]. It is often claimed that these targeted operating points can be used for different clinical purposes; for instance, a highly sensitive test is deemed effective at ruling out a disease when negative, whereas a highly specific test is effective at ruling in a disease when positive. However, these rules are misleading, as the diagnostic power of any test is determined by both its sensitivity and its specificity [62]. The tradeoff between specificity and sensitivity is explored in ROC analysis [63]. In this work, we analyze the usability of all the models for clinical settings, not only the individual operation points but across a range of values for the ability to predict a dichotomous outcome.

iii *Tangible improvement to physician's decision making.* Feature interpretation, as well as decision making logic, reliability, and robustness analysis of the machine learning models, is crucial and imperative for clinical applications. This task is much more complicated for recent techniques. Many recent efforts are short of explaining the decision-making logic and model interpretation in healthcare. In this work, we dive deeper into our machine learning model in an effort to further interpret the results, capabilities, and limitations. We investigate the most important factors that the machine learning model has learned in order to predict and classify an event. We review the clinical literature for additional verification and a better clinical understanding of the machine learning model. Finally, we examine the advantages and strengths of the proposed models.

6.2 LESSONS LEARNED

Throughout doing this dissertation, we have come across many challenges that made us rethink our approach. Some of the lessons we have learned about developing a human aging model and making augmented health intelligent systems:

- *Thinking about the user and setting:* most often, when training a machine learning model, we might have access to data modalities and features that are not easily accessible in the clinical practice. In some instances, certain measurements might be costly or time-consuming to collect. When designing and training the models, it is important to understand whether the user has access to the same input data with similar quality. How they will input the data and interact with the model. Fort instance, a PET scan might not be available to the physician, or the MRI might not have the same power or calibration parameters as the training data.

- *A closed-loop feedback system:* training data is never comprehensive and detailed enough to cover all scenarios, the human body's complexities, or population diversities. To overcome

that, the machine learning models need to continuously learn from the "human-in-the-loop" as well as additional data. When the model is uncertain about the prediction, it should have an option to receive feedback, maybe in the form of a nudge from a user, maybe falling back to manual mode. This interactive mechanism requires measurement of uncertainty in the prediction models, i.e., knowing what the model does not know, communicate that with the user, and enable a feedback mechanism.

- *Better data and more data:* in many use cases, more data usually enhances better algorithms. However, in healthcare, standardized and well-curated data is as important as more data. Most health data is available through EHR systems, which tends to have high error and subjectivity bias. Investment in standardized long-term data collection will pay off.

- *Data diversity matters:* demographic diversity in the age, gender, and racial/ethnic composition of data will have a major impact on healthcare disparities. For instance, sickle cell disease is more common in people of African, African American, or Mediterranean heritage [195]. Training a model on a population with European ancestry and using it in a hospital with a majority of African American patients would lead to hazardous health outcomes.

- *Nature is full of rare events:* many disorders have a very low prevalence, which would cause unbalanced class problems in machine learning. We should be diligent in addressing the rarity through both broader data collection as well as analytical solutions. Throughout the analysis, we should pay close attention to the ramifications of such rare events, making sure training and evaluation methods are adapted accordingly.

- *Data curation is often overlooked:* in practice, one-third of our project's time is spent on data curation, one-third on model training, and the rest on evaluation and interpretation. Data curation is often overlooked and considered as a tedious task. However, in healthcare-related projects, data curation helps the data scientist to understand the domain knowledge better and interact with the physicians. We often performed the curation task multiple times with the physicians to ensure data bias, leakage, imputation, etc., are handled properly. Sometimes data might be found inadequate due to significant missingness or data errors.

- *Qualities of a strong machine learning model for healthcare:* from the start, we need to ensure that the resulting models are reproducible, usable, and interpretable by the user. Also, generalizable to a targeted population.

- *Simple models are better than complex models:* a principle most widely known as Occam's razor, or the Law of Parsimony, states that "it is better, in explaining something, to use as few assumptions as possible" [196]. This idea is more formalized in statistical learning theory [197]; simply, among competing hypotheses that explain known observations equally well, one should choose the "simplest" one. A variation used in medicine is called the "Zebra": a doctor

should reject an exotic medical diagnosis when a more commonplace explanation is more likely [198]. In many healthcare use cases, interpretability and explainability of a machine learning model are more important than complex models even if they improve accuracy by a small percentage. For instance, in most cases, linear models can be interpreted and debugged more easily than neural nets. We can examine the weights assigned to each feature to figure out what (and how) is having the biggest impact in the model [199, 200].

- *Sometimes more complex models are needed:* often more complex models may not show improvements with simple features. However, with more complex features, it is more likely that we may require more complex models. Complex models are more successful with raw data captured from natural phenomena such as multi-omics, imaging, and speech data.

- *Unsupervised and semi-supervised more necessary than supervised learning:* in healthcare, most labels are collected by the physicians. The human judgment makes the data prone to errors and subjectivity bias. By relying on unsupervised and semi-supervised, we would be able to extract patterns and behaviors previously unknown to human.

- *Evaluation of unsupervised learning is hard:* still an open problem, especially in healthcare where we try to capture patterns previously unknown to human. Real-life clinical trials is a solution but a costly one.

- *Global data sharing is crucial:* considering the ethnicity and diversity complexities; the only way to developing highly generalizable solutions for public health benefits is global collaboration and data sharing.

- *Privacy and ethics matters:* obfuscation and anonymization of data is paramount to global data collection and sharing in healthcare. Similarly, deployment of machine learning models without ethical considerations will result in public distrust and pushback.

- *Open science, code sharing, and documentation:* we are facing a major reproducibility crisis in science [201]. More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. The reproducibility issue is no better in the Artificial Intelligence community. According to one study, only 6% of the papers presented at two top AI conferences in the past few years shared the algorithm's code [202]. To allow independent replication and extension of science, scientists must share their code and analytics notebooks publicly. This should include a readable and annotated step-by-step statistical and machine learning analysis. Many errors go unforeseen, public sharing, and review ensure that a minor data leakage would not jeopardize the whole system.

- *Domain knowledge matters:* much of a healthcare project is spent on problem formalization, interpreting the results, and evaluating their clinical impact. Computer scientists need to

gain adequate medical and health knowledge. They also need to have close collaboration with clinicians who are enthusiastic about data-driven and technology-based healthcare.

- *Multi-disciplinary teams:* from data collection to model test, and production deployment in a clinical setting, healthcare projects should comprise of experts in machine learning, computer systems, biostatistics, biology, medicine, ethics, and psychology. Building data science teams and integrating them into existing clinical workflows can be difficult and time-consuming. To maximize the accessibility and clinical impact of health data analytics, it is necessary to build multi-disciplinary teams that both transcend boundaries and is highly inclusive.

6.3 FUTURE WORK

This work is only a step to a much larger pursuit of healthy aging. There are several open problems and directions we will explore further in the future:

- *Coordination, collaboration, communication, for larger datasets:* plethora of worldwide collaborations are necessary to address challenges facing augmentation of the machine learning in healthcare. The challenges include but are not limited to generalizability, diversity, scale, standardization, and comprehensive evaluation. In the next decade, there need to be more data collection efforts similar to the UKBiobank [203], the NIH All of Us research programs [204] in terms of size, and International Parkinson's Disease Genomics Consortium (IPDGC) [205] in terms of international collaboration. Large studies address not only the generalizability, ethnic diversity, and under-represented population, but also rare events and disorders.

- *Privacy-preserving solutions:* large-scale data collection for the benefit of public health is only possible by ensuring individuals' privacy and providing adequate protection. These solutions will be a combination of law, policy, as well as technological solutions. As an example, the United States Genetic Information Nondiscrimination Act (GINA) [206], passed into law in 2008, prohibits discrimination by employers and health insurers based on genetic testing. The European Union General Data Protection Regulation (GDPR) [207], which became enforceable in 2018, regulates the protection of natural persons with regard to the processing of personal data and on the free movement of such data. These laws have provided individual protection but also difficulties for data sharing in science. Though still under research and development, technical solutions such as Federated Learning [208] and related decentralized approaches will help us enforce the protective policies while advancing the scientific discoveries.

- *Enhanced solutions for more data, more modalities, more longitudinal:* with larger datasets becoming available through the UKBiobank, the NIH All of Us research programs, and

multiple international efforts, we will have new opportunities for developing more comprehensive and generalizable models. At the same time, we are going to face newer engineering challenges. Instead of hundreds of individuals, we need to address millions of individual's multi-omics, imaging, biological, environmental, and geospatial data. With more longitudinal data collection starting from early adulthood, we would need techniques with much higher time-variant adaptability, which also addresses the data sparsity of longitudinal studies. We also need to explore novel solutions for handling rare events and class-imbalance problems.

- *Open science and democratization of tools:* one solution to address the health disparities and scientific reproducibility crisis is making data science tools more available and easier to use. For instance, there is a high barrier for junior scientists and non-biostatisticians to work on genomic data. To address this issue and making genomic and machine learning more accessible, we have been working on an automated machine learning tool for genomics called GenoML [209]. There is a need for more similar tools. Making these tools available through upcoming data science platforms such as Terra [210] will revolutionize open science in healthcare.

- *Robust and extensive clinical evaluation:* with more models developed by the community, we need standardized practices for evaluating the correctness, robustness, and generalizability of these models in clinical settings, across varying medical practices. This would require both policy and engineering solutions necessary to evaluate the performance and adverse effects of these solutions in the real-world.

- *Interpretability and Human-Computer Interaction:* as models used more and more in practice, we need to address the human interaction by making them more understandable for physicians as well as the patients. We need to recognize, understand, and engage with these users rigorously and systematically.

- *Broadening our perspective of complex disorders:* as we have learned, ALS, Alzheimer's, and Parkinson's are not single entities, but rather represent a collection of syndromes. There is a broad variability in the clinical manifestations of these complex disorders, which makes the diagnosis, prognosis, counseling, and clinical trial design limited [44]. Instead of analyzing individual disorders, we need to merge multiple disease datasets and let the data-driven methods deconstruct the heterogeneity within multiple cohorts. This will require close collaboration between areas of medicine, as well as a paradigm shift in our study designs and funding.

- *Broader healthcare impact:* human body is not just influenced by aging-related disorders. Many other life events can impact the aging trajectory. Events such as a traumatic accident or health epidemics can have a detrimental impact on aging. Expanding the models to other

areas of healthcare other than aging-related disorders would enable us to capture causal issues before adulthood or unrelated to aging.

- *Training multi-disciplinary scientists, the CS/ML+Health taskforce:* machine learning and computer science are contributing to the clinical understanding of the detection and treatment of health and aging-related disorders. However, the lack of a talented and skillful workforce is becoming a major challenge in deploying data science in practice. According to the LinkedIn August 2018 report, the demand for data science skills is rising across industries in the U.S. Consequently, the high demand has resulted in a countrywide shortage of 151,717 professionals with data science skills [211]. This shortage gap is much wider in healthcare, where talent is not only required to have data science expertise but also healthcare and biomedical knowledge and expertise. We believe that academia partnered with the government, and the private sector needs to play a national role in training the data science taskforce to empower analytics efforts in health and biomedical science.

REFERENCES

[1] J. R. Beard, A. Officer, I. A. De Carvalho, R. Sadana, A. M. Pot, J.-P. Michel, P. Lloyd-Sherlock, J. E. Epping-Jordan, G. G. Peeters, W. R. Mahanani et al., "The world report on ageing and health: a policy framework for healthy ageing," *The Lancet*, vol. 387, no. 10033, pp. 2145–2154, 2016.

[2] E. Craig, *The shorter Routledge encyclopedia of philosophy*. Routledge, 2005.

[3] C. Singer, *A short history of anatomy from the Greeks to Harvey*. Dover, 1957.

[4] J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny, *Descartes: Selected philosophical writings*. Cambridge: Cambridge University Press, 1988.

[5] B. Powell, "Descartes' machines," in *Proceedings of the Aristotelian Society*, vol. 71. JSTOR, 1970, pp. 209–222.

[6] M. Hawkins, ""A great and difficult thing": Understanding and explaining the human machine in restoration england," *Bodies/machines*, pp. 15–38, 2002.

[7] L. J. Kirmayer, "Mind and body as metaphors: hidden values in biomedicine," in *Biomedicine examined*. Springer, 1988, pp. 57–93.

[8] G. L. Engel, "The need for a new medical model: a challenge for biomedicine," *Science*, vol. 196, no. 4286, pp. 129–136, 1977.

[9] S. L. Aronoff, K. Berkowitz, B. Shreiner, and L. Want, "Glucose metabolism and regulation: beyond insulin and glucagon," *Diabetes spectrum*, vol. 17, no. 3, pp. 183–190, 2004.

[10] J. Fuller, "The new medical model: a renewed challenge for biomedicine," *CMAJ*, vol. 189, no. 17, pp. E640–E641, 2017.

[11] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England journal of medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[12] H. Yaghootkar, R. A. Scott, C. C. White, W. Zhang, E. Speliotes, P. B. Munroe, G. B. Ehret, J. C. Bis, C. S. Fox, M. Walker et al., "Genetic evidence for a normal-weight "metabolically obese" phenotype linking insulin resistance, hypertension, coronary artery disease, and type 2 diabetes," *Diabetes*, vol. 63, no. 12, pp. 4369–4377, 2014.

[13] J. C. Florez, "Precision medicine in diabetes: is it time?" *Diabetes care*, vol. 39, no. 7, pp. 1085–1088, 2016.

[14] A. M. Turing, "On computable numbers, with an application to the entscheidungsproblem," *Proceedings of the London mathematical society*, vol. 2, no. 1, pp. 230–265, 1937.

[15] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. October, pp. 433–60, 1950.

[16] D. Abramson, "Descartes' influence on turing," *Studies in History and Philosophy of Science Part A*, vol. 42, no. 4, pp. 544–551, 2011.

[17] T. G. Coleman, "A mathematical model of the human body in health, disease, and during treatment." *ISA transactions*, vol. 18, no. 3, pp. 65–73, 1979.

[18] T. G. Coleman, "Mathematical analysis of cardiovascular function," *IEEE transactions on biomedical engineering*, no. 4, pp. 289–294, 1985.

[19] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: promise and challenges," *Nature Reviews Cardiology*, vol. 13, no. 6, p. 350, 2016.

[20] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[21] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[22] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," *Bioinformatics*, vol. 32, no. 12, pp. 1832–1839, 2016.

[23] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[24] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[25] G. E. Box, "Science and statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.

[26] G. E. Box, "Robustness in the strategy of scientific model building," in *Robustness in statistics*. Elsevier, 1979, pp. 201–236.

[27] G. E. Box and N. R. Draper, *Empirical model-building and response surfaces*. John Wiley & Sons, 1987.

[28] S. C. Stearns, R. M. Nesse, D. R. Govindaraju, and P. T. Ellison, "Evolutionary perspectives on health and medicine," *Proceedings of the National Academy of Sciences*, vol. 107, no. suppl 1, pp. 1691–1695, 2010.

[29] R. M. Nesse and G. C. Williams, *Why we get sick: The new science of Darwinian medicine*. Vintage, 2012.

[30] J. von Neumann, "The mathematician," in *The Works of the Mind*. University of Chicago Press, 1947.

[31] S. Licher, S. K. Darweesh, F. J. Wolters, L. Fani, A. Heshmatollah, U. Mutlu, P. J. Koudstaal, J. Heeringa, M. J. Leening, M. K. Ikram et al., "Lifetime risk of common neurological diseases in the elderly population," *J Neurol Neurosurg Psychiatry*, vol. 90, no. 2, pp. 148–156, 2019.

[32] A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, A. M. Prina, B. Winblad, L. Jönsson, Z. Liu, and M. Prince, "The worldwide costs of dementia 2015 and comparisons with 2010," *Alzheimer's & Dementia*, vol. 13, no. 1, pp. 1–7, 2017.

[33] V. L. Feigin, B. Norrving, and G. A. Mensah, "Global burden of stroke," *Circulation research*, vol. 120, no. 3, pp. 439–448, 2017.

[34] S. L. Kowal, T. M. Dall, R. Chakrabarti, M. V. Storm, and A. Jain, "The current and projected economic burden of parkinson's disease in the united states," *Movement Disorders*, vol. 28, no. 3, pp. 311–318, 2013.

[35] V. L. Feigin, A. A. Abajobir, K. H. Abate, F. Abd-Allah, A. M. Abdulle, S. F. Abera, G. Y. Abyu, M. B. Ahmed, A. N. Aichour, I. Aichour et al., "Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet Neurology*, vol. 16, no. 11, pp. 877–897, 2017.

[36] S. Norton, F. E. Matthews, D. E. Barnes, K. Yaffe, and C. Brayne, "Potential for primary prevention of alzheimer's disease: an analysis of population-based data," *The Lancet Neurology*, vol. 13, no. 8, pp. 788–794, 2014.

[37] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: astronomical or genomical?" *PLoS biology*, vol. 13, no. 7, p. e1002195, 2015.

[38] M. A. Nalls, C. Blauwendraat, C. L. Vallerga, K. Heilbron, S. Bandres-Ciga, D. Chang, M. Tan, D. A. Kia, A. J. Noyce, A. Xue et al., "Identification of novel risk loci, causal insights, and heritable risk for parkinson's disease: a meta-analysis of genome-wide association studies," *The Lancet Neurology*, vol. 18, no. 12, pp. 1091–1102, 2019.

[39] H. Iwaki, C. Blauwendraat, H. L. Leonard, J. J. Kim, G. Liu, J. Maple-Grødem, J.-C. Corvol, L. Pihlstrøm, M. van Nimwegen, S. J. Hutten et al., "Genomewide association study of parkinson's disease clinical biomarkers in 12 longitudinal patients' cohorts," *Movement Disorders*, 2019.

[40] H. Iwaki, C. Blauwendraat, H. L. Leonard, G. Liu, J. Maple-Grødem, J.-C. Corvol, L. Pihlstrøm, M. van Nimwegen, S. J. Hutten, K.-D. H. Nguyen et al., "Genetic risk of parkinson disease and progression:: An analysis of 13 longitudinal cohorts," *Neurology Genetics*, vol. 5, no. 4, p. e348, 2019.

[41] H. Iwaki, C. Blauwendraat, M. B. Makarious, S. Bandres-Ciga, H. L. Leonard, J. R. Gibbs, D. G. Hernandez, S. W. Scholz, F. Faghri, M. A. Nalls et al., "Penetrance of parkinson's disease in lrrk2 p. g2019s carriers is modified by a polygenic risk score," *BioRxiv*, p. 738260, 2019.

[42] M. A. Nalls, C. Blauwendraat, C. L. Vallerga, K. Heilbron, S. Bandres-Ciga, D. Chang, M. Tan, D. A. Kia, A. J. Noyce, A. Xue et al., "Expanding parkinsons disease genetics: novel risk loci, genomic context, causal insights and heritable risk," *BioRxiv*, p. 388165, 2019.

[43] F. Faghri, S. H. Hashemi, H. Leonard, S. W. Scholz, R. H. Campbell, M. A. Nalls, and A. B. Singleton, "Predicting onset, progression, and clinical subtypes of parkinson disease using machine learning," *bioRxiv*, p. 338913, 2018.

[44] H. Leonard, C. Blauwendraat, L. Krohn, F. Faghri, H. Iwaki, G. Ferguson, A. G. Day-Williams, D. J. Stone, A. B. Singleton, M. A. Nalls et al., "Genetic variability and potential effects on clinical trial outcomes: perspectives in parkinsons disease," *Journal of Medical Genetics*, 2019.

[45] M. A. Nalls, C. Blauwendraat, C. L. Vallerga, K. Heilbron, S. Bandres-Ciga, D. Chang, M. Tan, D. A. Kia, A. J. Noyce, A. Xue et al., "Parkinsons disease genetics: Identifying novel risk loci, providing causal insights and improving estimates of heritable risk," *BioRxiv*, p. 388165, 2018.

[46] D. Chang, M. A. Nalls, I. B. Hallgrímsdóttir, J. Hunkapiller, M. van der Brug, F. Cai, G. A. Kerchner, G. Ayalon, B. Bingol, M. Sheng et al., "A meta-analysis of genome-wide association studies identifies 17 new parkinson's disease risk loci," *Nature genetics*, vol. 49, no. 10, p. 1511, 2017.

[47] S. Bandres-Ciga, S. Saez-Atienzar, L. Bonet-Ponce, K. Billingsley, D. Vitale, C. Blauwendraat, J. R. Gibbs, L. Pihlstrøm, Z. Gan-Or, I. P. D. G. C. (IPDGC) et al., "The endocytic membrane trafficking pathway plays a major role in the risk of parkinson's disease," *Movement Disorders*, vol. 34, no. 4, pp. 460–468, 2019.

[48] S. Bandres-Ciga and F. Faghri, "Unraveling the genetic complexity of alzheimer disease with mendelian randomization," 2019.

[49] V. K. Satone, R. Kaur, H. Leonard, H. Iwaki, L. Sargent, S. W. Scholz, M. A. Nalls, A. B. Singleton, F. Faghri, R. H. Campbell et al., "Predicting alzheimers disease progression trajectory and clinical subtypes using machine learning," *bioRxiv*, p. 792432, 2019.

[50] V. Satone, R. Kaur, F. Faghri, M. A. Nalls, A. B. Singleton, and R. H. Campbell, "Learning the progression and clinical subtypes of alzheimer's disease from longitudinal clinical data," *NeurIPS 2018 Workshop on Machine Learning for Health (ML4H)*, 2018.

[51] C. Blauwendraat, O. Pletnikova, J. T. Geiger, N. A. Murphy, Y. Abramzon, G. Rudow, A. Mamais, M. S. Sabir, B. Crain, S. Ahmed et al., "Genetic analysis of neurodegenerative diseases in a pathology cohort," *Neurobiology of aging*, vol. 76, pp. 214–e1, 2019.

[52] J. O. Johnson, R. Chia, R. H. Brown Jr, and J. E. Landers, "Mutations in the sptlc1 gene are a cause of amyotrophic lateral sclerosis that may be amenable to serine supplementation," *bioRxiv*, 2019.

[53] S. Bandres-Ciga, A. J. Noyce, G. Hemani, A. Nicolas, A. Calvo, G. Mora, I. Consortium, A. Arosio, M. Barberis, I. Bartolomei et al., "Shared polygenic risk and causal inferences in amyotrophic lateral sclerosis," *Annals of neurology*, vol. 85, no. 4, pp. 470–481, 2019.

[54] A. Nicolas, K. P. Kenna, A. E. Renton, N. Ticozzi, F. Faghri, R. Chia, J. A. Dominov, B. J. Kenna, M. A. Nalls, P. Keagle et al., "Genome-wide analyses identify kif5a as a novel als gene," *Neuron*, vol. 97, no. 6, pp. 1268–1283, 2018.

[55] Y.-W. Lin, Y. Zhou, F. Faghri, M. J. Shaw, and R. H. Campbell, "Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory," *PloS one*, vol. 14, no. 7, p. e0218942, 2019.

[56] F. Faghri, S. Bazarbayev, M. Overholt, R. Farivar, R. H. Campbell, and W. H. Sanders, "Failure scenario as a service (FSaaS) for hadoop clusters," in *Proceedings of the Workshop on Secure and Dependable Middleware for Cloud Monitoring and Management*. ACM, 2012, p. 5.

[57] F. Faghri, S. H. Hashemi, M. Babaeizadeh, M. A. Nalls, S. Sinha, and R. H. Campbell, "Toward scalable machine learning and data mining: the bioinformatics case," *arXiv preprint arXiv:1710.00112*, 2017.

[58] S. H. Hashemi, F. Faghri, P. Rausch, and R. H. Campbell, "World of empowered IoT users," in *2016 IEEE First International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 2016, pp. 13–24.

[59] S. H. Hashemi, F. Faghri, and R. H. Campbell, "Decentralized user-centric access control using pubsub over blockchain," *arXiv preprint arXiv:1710.00110*, 2017.

[60] C. Blauwendraat, F. Faghri, L. Pihlstrom, J. T. Geiger, A. Elbaz, S. Lesage, J.-C. Corvol, P. May, A. Nicolas, Y. Abramzon et al., "NeuroChip, an updated version of the neurox genotyping platform to rapidly screen for variants associated with neurological diseases," *Neurobiology of aging*, vol. 57, pp. 247–e9, 2017.

[61] D. G. Altman and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity." *BMJ: British Medical Journal*, vol. 308, no. 6943, p. 1552, 1994.

[62] D. Pewsner, M. Battaglia, C. Minder, A. Marx, H. C. Bucher, and M. Egger, "Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution," *Bmj*, vol. 329, no. 7459, pp. 209–213, 2004.

[63] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[64] A. J. Hughes, S. E. Daniel, L. Kilford, and A. J. Lees, "Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 55, no. 3, pp. 181–184, 1992.

[65] R. B. Postuma, D. Berg, M. Stern, W. Poewe, C. W. Olanow, W. Oertel, J. Obeso, K. Marek, I. Litvan, A. E. Lang et al., "Mds clinical diagnostic criteria for parkinson's disease," *Movement Disorders*, vol. 30, no. 12, pp. 1591–1601, 2015.

[66] G. T. Stebbins, C. G. Goetz, D. J. Burn, J. Jankovic, T. K. Khoo, and B. C. Tilley, "How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified parkinson's disease rating scale: comparison with the unified parkinson's disease rating scale," *Movement Disorders*, vol. 28, no. 5, pp. 668–670, 2013.

[67] J. Jankovic, M. McDermott, J. Carter, S. Gauthier, C. Goetz, L. Golbe, S. Huber, W. Koller, C. Olanow, I. Shoulson et al., "Variable expression of parkinson's disease: A base-line analysis of the dat atop cohort," *Neurology*, vol. 40, no. 10, pp. 1529–1529, 1990.

[68] W. J. Zetusky, J. Jankovic, and F. J. Pirozzolo, "The heterogeneity of parkinson's disease: clinical and prognostic implications," *Neurology*, vol. 35, no. 4, pp. 522–522, 1985.

[69] S. M. van Rooden, W. J. Heiser, J. N. Kok, D. Verbaan, J. J. van Hilten, and J. Marinus, "The identification of parkinson's disease subtypes using cluster analysis: a systematic review," *Movement disorders*, vol. 25, no. 8, pp. 969–978, 2010.

[70] S.-M. Fereshtehnejad, S. R. Romenets, J. B. Anang, V. Latreille, J.-F. Gagnon, and R. B. Postuma, "New clinical subtypes of parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes," *JAMA neurology*, vol. 72, no. 8, pp. 863–873, 2015.

[71] S.-M. Fereshtehnejad, Y. Zeighami, A. Dagher, and R. B. Postuma, "Clinical criteria for subtyping parkinson's disease: biomarkers and longitudinal progression," *Brain*, vol. 140, no. 7, pp. 1959–1976, 2017.

[72] M. A. Nalls, C. Y. McLean, J. Rick, S. Eberly, S. J. Hutten, K. Gwinn, M. Sutherland, M. Martinez, P. Heutink, N. M. Williams et al., "Diagnosis of parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study," *The Lancet Neurology*, vol. 14, no. 10, pp. 1002–1009, 2015.

[73] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel et al., "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results," *Movement disorders: official journal of the Movement Disorder Society*, vol. 23, no. 15, pp. 2129–2170, 2008.

[74] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[75] J. Brandt, "The hopkins verbal learning test: Development of a new memory test with six equivalent forms," *The Clinical Neuropsychologist*, vol. 5, no. 2, pp. 125–142, 1991.

[76] H. Goodglass, E. Kaplan, and B. Barresi, *The assessment of aphasia and related disorders*. Lippincott Williams & Wilkins, 2001.

[77] D. Wechsler, *WAiS-iii*. Psychological Corporation San Antonio, TX, 1997.

[78] A. L. Benton, N. R. Varney, and K. d. Hamsher, "Visuospatial judgment: A clinical test," *Archives of neurology*, vol. 35, no. 6, pp. 364–367, 1978.

[79] A. Smith, *Symbol digit modalities test*. Western Psychological Services Los Angeles, CA, 1982.

[80] M. Visser, J. Marinus, A. M. Stiggelbout, and J. J. Van Hilten, "Assessment of autonomic dysfunction in parkinson's disease: the scopa-aut," *Movement disorders: official journal of the Movement Disorder Society*, vol. 19, no. 11, pp. 1306–1312, 2004.

[81] C. D. Spielberger and R. L. Gorsuch, *State-trait anxiety inventory for adults: sampler set: manual, test, scoring key*. Mind Garden, 1983.

[82] J. I. Sheikh and J. A. Yesavage, "Geriatric depression scale (gds): recent evidence and development of a shorter version." *Clinical Gerontologist: The Journal of Aging and Mental Health*, 1986.

[83] D. Weintraub, S. Hoops, J. A. Shea, K. E. Lyons, R. Pahwa, E. D. Driver-Dunckley, C. H. Adler, M. N. Potenza, J. Miyasaki, A. D. Siderowf et al., "Validation of the questionnaire for impulsive-compulsive disorders in parkinson's disease," *Movement disorders: official journal of the Movement Disorder Society*, vol. 24, no. 10, pp. 1461–1467, 2009.

[84] K. Stiasny-Kolster, G. Mayer, S. Schäfer, J. C. Möller, M. Heinzel-Gutenbrunner, and W. H. Oertel, "The rem sleep behavior disorder screening questionnairea new diagnostic instrument," *Movement disorders*, vol. 22, no. 16, pp. 2386–2393, 2007.

[85] M. W. Johns, "A new method for measuring daytime sleepiness: the epworth sleepiness scale," *sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[86] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[87] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[88] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*. M. Dekker New York, 1988, vol. 84.

[89] R. L. Prentice, "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika*, vol. 73, no. 1, pp. 1–11, 1986.

[90] G. Schwarz et al., "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

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

[92] M. M. Corrada, R. Brookmeyer, A. Paganini-Hill, D. Berlau, and C. H. Kawas, "Dementia incidence continues to increase with age in the oldest old: the 90+ study," *Annals of neurology*, vol. 67, no. 1, pp. 114–121, 2010.

[93] C. P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang et al., "Global prevalence of dementia: a delphi consensus study," *The lancet*, vol. 366, no. 9503, pp. 2112–2117, 2005.

[94] V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, O. Salvado, C. Szoeke, S. L. Macaulay, R. Martins, P. Maruff et al., "Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study," *The Lancet Neurology*, vol. 12, no. 4, pp. 357–367, 2013.

[95] C. R. Jack Jr, V. J. Lowe, S. D. Weigand, H. J. Wiste, M. L. Senjem, D. S. Knopman, M. M. Shiung, J. L. Gunter, B. F. Boeve, B. J. Kemp et al., "Serial pib and mri in normal, mild cognitive impairment and alzheimer's disease: implications for sequence of pathological events in alzheimer's disease," *Brain*, vol. 132, no. 5, pp. 1355–1365, 2009.

[96] E. M. Reiman, Y. T. Quiroz, A. S. Fleisher, K. Chen, C. Velez-Pardo, M. Jimenez-Del-Rio, A. M. Fagan, A. R. Shah, S. Alvarez, A. Arbelaez et al., "Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant alzheimer's disease in the presenilin 1 e280a kindred: a case-control study," *The Lancet Neurology*, vol. 11, no. 12, pp. 1048–1056, 2012.

[97] R. J. Bateman, C. Xiong, T. L. Benzinger, A. M. Fagan, A. Goate, N. C. Fox, D. S. Marcus, N. J. Cairns, X. Xie, T. M. Blazey et al., "Clinical and biomarker changes in dominantly inherited alzheimer's disease," *New England Journal of Medicine*, vol. 367, no. 9, pp. 795–804, 2012.

[98] H. Braak, D. R. Thal, E. Ghebremedhin, and K. Del Tredici, "Stages of the pathologic process in alzheimer disease: age categories from 1 to 100 years," *Journal of Neuropathology & Experimental Neurology*, vol. 70, no. 11, pp. 960–969, 2011.

[99] Alzheimer's Association, "2019 alzheimer's disease facts and figures report," 2019. [Online]. Available: https://www.alz.org/media/documents/alzheimers-facts-and-figures-2019-r.pdf

[100] Alzheimer's Association, "Medications for memory loss," 2019. [Online]. Available: https://alz.org/alzheimers-dementia/treatments/medications-for-memory

[101] Alzheimer's Association, "Treatments for alzheimer's and dementia," 2019. [Online]. Available: https://alz.org/alzheimers-dementia/treatments

[102] Alzheimer's Association, "Alzheimer's and dementia stages," 2019. [Online]. Available: https://alz.org/alzheimers-dementia/stages

[103] D. M. Holtzman, J. C. Morris, and A. M. Goate, "Alzheimer's disease: the challenge of the second century," *Science translational medicine*, vol. 3, no. 77, pp. 77sr1–77sr1, 2011.

[104] P. Boyle, R. Wilson, N. Aggarwal, Y. Tang, and D. Bennett, "Mild cognitive impairment: risk of alzheimer disease and rate of cognitive decline," *Neurology*, vol. 67, no. 3, pp. 441–445, 2006.

[105] O. Hansson, H. Zetterberg, P. Buchhave, E. Londos, K. Blennow, and L. Minthon, "Association between csf biomarkers and incipient alzheimer's disease in patients with mild cognitive impairment: a follow-up study," *The Lancet Neurology*, vol. 5, no. 3, pp. 228–234, 2006.

[106] S. A. Hassan and T. Khan, "A machine learning model to predict the onset of alzheimer disease using potential cerebrospinal fluid (csf) biomarkers," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, pp. 124–131, 2017.

[107] J. Moreland, T. Urhemaa, M. Van Gils, J. Lötjönen, J. Wolber, and C. J. Buckley, "Validation of prognostic biomarker scores for predicting progression of dementia in patients with amnestic mild cognitive impairment," *Nuclear medicine communications*, vol. 39, no. 4, p. 297, 2018.

[108] T. Altaf, S. M. Anwar, N. Gul, M. N. Majeed, and M. Majid, "Multi-class alzheimer's disease classification using image and clinical features," *Biomedical Signal Processing and Control*, vol. 43, pp. 64–74, 2018.

[109] "Alzheimer's Disease Neuroimaging Initiative: ADNI," 2016. [Online]. Available: http://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/clinical/ADNI3_Protocol.pdf

[110] L. Berg, "Clinical dementia rating," *The British Journal of Psychiatry*, vol. 145, no. 3, pp. 339–339, 1984.

[111] S. Zaidi, M. G. Kat, and J. F. de Jonghe, "Clinician and caregiver agreement on neuropsychiatric symptom severity: a study using the neuropsychiatric inventory–clinician rating scale (npi-c)," *International psychogeriatrics*, vol. 26, no. 7, pp. 1139–1145, 2014.

[112] C. B. Dodrill, "A neuropsychological battery for epilepsy," *Epilepsia*, vol. 19, no. 6, pp. 611–623, 1978.

[113] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[114] J. C. Allaire and M. Marsiske, "Everyday cognition: age and intellectual ability correlates." *Psychology and aging*, vol. 14, no. 4, p. 627, 1999.

[115] G. Salomon, D. N. Perkins, and T. Globerson, "Partners in cognition: Extending human intelligence with intelligent technologies," *Educational researcher*, vol. 20, no. 3, pp. 2–9, 1991.

[116] G. G. Fillenbaum and M. A. Smyer, "The development, validity, and reliability of the oars multidimensional functional assessment questionnaire," *Journal of gerontology*, vol. 36, no. 4, pp. 428–434, 1981.

[117] W. J. Strittmatter, K. H. Weisgraber, D. Y. Huang, L.-M. Dong, G. S. Salvesen, M. Pericak-Vance, D. Schmechel, A. M. Saunders, D. Goldgaber, and A. D. Roses, "Binding of human apolipoprotein e to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset alzheimer disease," *Proceedings of the National Academy of Sciences*, vol. 90, no. 17, pp. 8098–8102, 1993.

[118] J. Raber, D. Wong, G.-Q. Yu, M. Buttini, R. W. Mahley, R. E. Pitas, and L. Mucke, "Alzheimer's disease: Apolipoprotein e and cognitive performance," *Nature*, vol. 404, no. 6776, p. 352, 2000.

[119] Y. Stern, S. Albert, M.-X. Tang, and W.-Y. Tsai, "Rate of memory decline in ad is related to education and occupation: cognitive reserve?" *Neurology*, vol. 53, no. 9, pp. 1942–1942, 1999.

[120] J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, G. W. Beecham et al., "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature genetics*, vol. 45, no. 12, p. 1452, 2013.

[121] I. O. Korolev, L. L. Symonds, A. C. Bozoki, A. D. N. Initiative et al., "Predicting progression from mild cognitive impairment to alzheimer's dementia using clinical, mri, and plasma biomarkers via probabilistic pattern classification," *PloS one*, vol. 11, no. 2, p. e0138866, 2016.

[122] T. G. Beach, S. E. Monsell, L. E. Phillips, and W. Kukull, "Accuracy of the clinical diagnosis of alzheimer disease at national institute on aging alzheimer disease centers, 2005–2010," *Journal of neuropathology and experimental neurology*, vol. 71, no. 4, pp. 266–273, 2012.

[123] National Institutes on Aging / National Institutes of Health, "Accelerating medicines partnership - alzheimer's disease (amp-ad)," 2019. [Online]. Available: https://www.nia.nih.gov/research/amp-ad

[124] K. C. Arthur, A. Calvo, T. R. Price, J. T. Geiger, A. Chio, and B. J. Traynor, "Projected increase in amyotrophic lateral sclerosis from 2015 to 2040," *Nature communications*, vol. 7, p. 12408, 2016.

[125] D. Hirtz, D. Thurman, K. Gwinn-Hardy, M. Mohamed, A. Chaudhuri, and R. Zalutsky, "How common are the "common" neurologic disorders?" *Neurology*, vol. 68, no. 5, pp. 326–337, 2007.

[126] R. Chia, A. Chiò, and B. J. Traynor, "Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications," *The Lancet Neurology*, vol. 17, no. 1, pp. 94–102, 2018.

[127] M. R. Turner, O. Hardiman, M. Benatar, B. R. Brooks, A. Chio, M. De Carvalho, P. G. Ince, C. Lin, R. G. Miller, H. Mitsumoto et al., "Controversies and priorities in amyotrophic lateral sclerosis," *The Lancet Neurology*, vol. 12, no. 3, pp. 310–322, 2013.

[128] S. Byrne, P. Bede, M. Elamin, K. Kenna, C. Lynch, R. McLaughlin, and O. Hardiman, "Proposed criteria for familial amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis*, vol. 12, no. 3, pp. 157–159, 2011.

[129] A. Chiò, A. Calvo, C. Moglia, L. Mazzini, G. Mora et al., "Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 82, no. 7, pp. 740–746, 2011.

[130] N. Geevasinga, J. Howells, P. Menon, M. van den Bos, K. Shibuya, J. M. Matamala, S. B. Park, K. Byth, M. C. Kiernan, and S. Vucic, "Amyotrophic lateral sclerosis diagnostic index: Toward a personalized diagnosis of als," *Neurology*, vol. 92, no. 6, pp. e536–e547, 2019.

[131] M. de Carvalho, R. Dengler, A. Eisen, J. D. England, R. Kaji, J. Kimura, K. Mills, H. Mitsumoto, H. Nodera, J. Shefner et al., "Electrodiagnostic criteria for diagnosis of als," *Clinical neurophysiology*, vol. 119, no. 3, pp. 497–503, 2008.

[132] B. R. Brooks, "El escorial world federation of neurology criteria for the diagnosis of amyotrophic lateral sclerosis," *Journal of the neurological sciences*, vol. 124, pp. 96–107, 1994.

[133] Piemonte and Valle d'Aosta Register for Amyotrophic Lateral Sclerosis (PARALS), "Incidence of als in italy: evidence for a uniform frequency in western countries," *Neurology*, vol. 56, no. 2, pp. 239–244, 2001.

[134] J. Mandrioli, S. Biguzzi, C. Guidi, E. Venturini, E. Sette, E. Terlizzi, A. Ravasio, M. Casmiro, F. Salvi, R. Liguori et al., "Epidemiology of amyotrophic lateral sclerosis in emilia romagna region (italy): A population based study," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 15, no. 3-4, pp. 262–268, 2014.

[135] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: a critical evaluation," *BMC medical informatics and decision making*, vol. 16, no. 3, p. 74, 2016.

[136] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[137] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[138] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[139] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[140] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

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

[142] Centers for Medicare & Medicaid Services, "Fy 2018 hospital inpatient pps final rule," 2018. [Online]. Available: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2018-IPPS-Final-Rule-Home-Page-Items/FY2018-IPPS-Final-Rule-Data-Files.html

[143] A. A. Kramer, T. L. Higgins, and J. E. Zimmerman, "The association between icu readmission rate and patient outcomes," *Critical care medicine*, vol. 41, no. 1, pp. 24–33, 2013.

[144] C. R. Ponzoni, T. D. Corrêa, R. R. Filho, A. Serpa Neto, M. S. Assunção, A. Pardini, and G. P. Schettino, "Readmission to the intensive care unit: incidence, risk factors, resource use, and outcomes. a retrospective cohort study," *Annals of the American Thoracic Society*, vol. 14, no. 8, pp. 1312–1319, 2017.

[145] T. Desautels, R. Das, J. Calvert, M. Trivedi, C. Summers, D. J. Wales, and A. Ercole, "Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach," *BMJ open*, vol. 7, no. 9, p. e017199, 2017.

[146] L. M. Chen, C. M. Martin, S. P. Keenan, and W. J. Sibbald, "Patients readmitted to the intensive care unit during the same hospitalization: clinical features and outcomes," *Critical care medicine*, vol. 26, no. 11, pp. 1834–1841, 1998.

[147] H. B. Rubins and M. A. Moskowitz, "Discharge decision-making in a medical intensive care unit: identifying patients at high risk of unexpected death or unit readmission," *The American journal of medicine*, vol. 84, no. 5, pp. 863–869, 1988.

[148] D. E. Singer, A. G. Mulley, G. E. Thibault, and G. O. Barnett, "Unexpected readmissions to the coronary-care unit during recovery from acute myocardial infarction," *New England Journal of Medicine*, vol. 304, no. 11, pp. 625–629, 1981.

[149] C. A. Baillie, C. VanZandbergen, G. Tait, A. Hanish, B. Leas, B. French, C. William Hanson, M. Behta, and C. A. Umscheid, "The readmission risk flag: Using the electronic health record to automatically identify patients at risk for 30-day readmission," *Journal of hospital medicine*, vol. 8, no. 12, pp. 689–695, 2013.

[150] Y. Choi, C. Y.-I. Chiu, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 41, 2016.

[151] E. Shadmi, N. Flaks-Manov, M. Hoshen, O. Goldman, H. Bitterman, and R. D. Balicer, "Predicting 30-day readmissions with preadmission electronic health record data," *Medical care*, vol. 53, no. 3, pp. 283–289, 2015.

[152] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless, "Mining high-dimensional administrative claims data to predict early hospital readmissions," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 272–279, 2014.

[153] M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, and E. Liu, "Predicting all-cause risk of 30-day hospital readmission using artificial neural networks," *PloS one*, vol. 12, no. 7, p. e0181173, 2017.

[154] 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.

[155] I. Shams, S. Ajorlou, and K. Yang, "A predictive analytics approach to reducing avoidable hospital readmission," *arXiv preprint arXiv:1402.5991*, 2014.

[156] A. E. Nijhawan, E. Kitchell, S. S. Etherton, P. Duarte, E. A. Halm, and M. K. Jain, "Half of 30-day hospital readmissions among hiv-infected patients are potentially preventable," *AIDS patient care and STDs*, vol. 29, no. 9, pp. 465–473, 2015.

[157] H. Kim, J. S. Ross, G. D. Melkus, Z. Zhao, and K. Boockvar, "Scheduled and unscheduled hospital readmissions among diabetes patients," *The American journal of managed care*, vol. 16, no. 10, p. 760, 2010.

[158] M. A. McAdams-DeMarco, A. Law, M. L. Salter, E. Chow, M. Grams, J. Walston, and D. L. Segev, "Frailty and early hospital readmission after kidney transplantation," *American journal of transplantation*, vol. 13, no. 8, pp. 2091–2095, 2013.

[159] S. Curto, J. P. Carvalho, C. Salgado, S. M. Vieira, and J. M. Sousa, "Predicting icu readmissions based on bedside medical text notes," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2016, pp. 2144–a.

[160] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. V. Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *arXiv preprint arXiv:1703.07771*, 2017.

[161] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Machine Learning for Healthcare Conference*, 2016, pp. 73–100.

[162] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[163] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1495–1504.

[164] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[165] S. E. Brown, S. J. Ratcliffe, and S. D. Halpern, "An empirical derivation of the optimal time interval for defining icu readmissions," *Medical care*, vol. 51, no. 8, p. 706, 2013.

[166] F. S. Hosein, N. Bobrovitz, S. Berthelot, D. Zygun, W. A. Ghali, and H. T. Stelfox, "A systematic review of tools for predicting severe adverse events following patient discharge from intensive care units," *Critical Care*, vol. 17, no. 3, p. R102, 2013.

[167] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration," *Journal of biomedical informatics*, vol. 53, pp. 220–228, 2015.

[168] C. E. Kennedy and J. P. Turley, "Time series analysis as input for clinical predictive modeling: Modeling cardiac arrest in a pediatric icu," *Theoretical Biology and Medical Modelling*, vol. 8, no. 1, p. 40, 2011.

[169] J. Lee and R. G. Mark, "An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care," *Biomedical engineering online*, vol. 9, no. 1, p. 62, 2010.

[170] C. Hug, "Detecting hazardous intensive care patient episodes using real-time mortality models," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[171] M. Hoogendoorn, A. El Hassouni, K. Mok, M. Ghassemi, and P. Szolovits, "Prediction using patient comparison vs. modeling: A case study for mortality prediction," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 2464–2467.

[172] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.

[173] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[174] Q. Mao, M. Jay, J. L. Hoffman, J. Calvert, C. Barton, D. Shimabukuro, L. Shieh, U. Chettipally, G. Fletcher, Y. Kerem et al., "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu," *BMJ open*, vol. 8, no. 1, p. e017833, 2018.

[175] M. S. Pepe, H. Janes, C. I. Li, P. M. Bossuyt, Z. Feng, and J. Hilden, "Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility?" *Clinical chemistry*, vol. 62, no. 5, pp. 737–742, 2016.

[176] B. Wellner, J. Grand, E. Canzone, M. Coarr, P. W. Brady, J. Simmons, E. Kirkendall, N. Dean, M. Kleinman, and P. Sylvester, "Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements," *JMIR medical informatics*, vol. 5, no. 4, p. e45, 2017.

[177] J. Calvert, J. Hoffman, C. Barton, D. Shimabukuro, M. Ries, U. Chettipally, Y. Kerem, M. Jay, S. Mataraso, and R. Das, "Cost and mortality impact of an algorithm-driven sepsis prediction system," *Journal of medical economics*, vol. 20, no. 6, pp. 646–651, 2017.

[178] A. C. Alba, T. Agoritsas, M. Walsh, S. Hanna, A. Iorio, P. Devereaux, T. McGinn, and G. Guyatt, "Discrimination and calibration of clinical prediction models: users guides to the medical literature," *Jama*, vol. 318, no. 14, pp. 1377–1384, 2017.

[179] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian journal of ophthalmology*, vol. 56, no. 1, p. 45, 2008.

[180] F. Casalini, S. Salvetti, S. Memmini, E. Lucaccini, G. Massimetti, P. L. Lopalco, and G. P. Privitera, "Unplanned readmissions within 30 days after discharge: improving quality through easy prediction," *International Journal for Quality in Health Care*, vol. 29, no. 2, pp. 256–261, 2017.

[181] M. Zhang, C. D. J. Holman, S. D. Price, F. M. Sanfilippo, D. B. Preen, and M. K. Bulsara, "Comorbidity and repeat admission to hospital for adverse drug reactions in older adults: retrospective cohort study," *Bmj*, vol. 338, p. a2752, 2009.

[182] C. Berry, M. Brett, K. Stevenson, J. McMurray, and J. Norrie, "Nature and prognostic importance of abnormal glucose tolerance and diabetes in acute heart failure," *Heart*, vol. 94, no. 3, pp. 296–304, 2008.

[183] N. Evans and K. Dhatariya, "Assessing the relationship between admission glucose levels, subsequent length of hospital stay, readmission and mortality," *Clinical Medicine*, vol. 12, no. 2, pp. 137–139, 2012.

[184] K. M. Dungan, "The effect of diabetes on hospital readmissions," 2012.

[185] M. Emons, J. Bae, B. Hoogwerf, S. Kindermann, R. Taylor, and B. Nathanson, "Risk factors for 30-day readmission following hypoglycemia-related emergency room and inpatient admissions," *BMJ Open Diabetes Research and Care*, vol. 4, no. 1, p. e000160, 2016.

[186] J. M. Vinson, M. W. Rich, J. C. Sperry, A. S. Shah, and T. McNamara, "Early readmission of elderly patients with congestive heart failure," *Journal of the American Geriatrics Society*, vol. 38, no. 12, pp. 1290–1295, 1990.

[187] P. S. Keenan, S.-L. T. Normand, Z. Lin, E. E. Drye, K. R. Bhat, J. S. Ross, J. D. Schuur, B. D. Stauffer, S. M. Bernheim, A. J. Epstein et al., "An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure," *Circulation: Cardiovascular Quality and Outcomes*, vol. 1, no. 1, pp. 29–37, 2008.

[188] B. G. Hammill, L. H. Curtis, G. C. Fonarow, P. A. Heidenreich, C. W. Yancy, E. D. Peterson, and A. F. Hernandez, "Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization," *Circulation: Cardiovascular Quality and Outcomes*, vol. 4, no. 1, pp. 60–67, 2011.

[189] A. T. Mathew, G. F. Strippoli, M. Ruospo, and S. Fishbane, "Reducing hospital readmissions in patients with end-stage kidney disease," *Kidney international*, vol. 88, no. 6, pp. 1250–1260, 2015.

[190] R. Zager and R. Altschuld, "Body temperature: an important determinant of severity of ischemic renal injury," *American Journal of Physiology-Renal Physiology*, vol. 251, no. 1, pp. F87–F93, 1986.

[191] M. M. Sood, D. Roberts, P. Komenda, J. Bueti, M. Reslerova, J. Mojica, and C. Rigatto, "End-stage renal disease status and critical illness in the elderly," *Clinical Journal of the American Society of Nephrology*, vol. 6, no. 3, pp. 613–619, 2011.

[192] I. De Alba and A. Amin, "Pneumonia readmissions: risk factors and implications," *Ochsner Journal*, vol. 14, no. 4, pp. 649–654, 2014.

[193] E. A. Halm, M. J. Fine, W. N. Kapoor, D. E. Singer, T. J. Marrie, and A. L. Siu, "Instability on hospital discharge and the risk of adverse outcomes in patients with pneumonia," *Archives of internal medicine*, vol. 162, no. 11, pp. 1278–1284, 2002.

[194] T. Guo, Z. Xu, X. Yao, H. Chen, K. Aberer, and K. Funaya, "Robust online time series prediction with recurrent neural networks," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Ieee, 2016, pp. 816–825.

[195] K. L. Hassell, "Population estimates of sickle cell disease in the us," *American journal of preventive medicine*, vol. 38, no. 4, pp. S512–S521, 2010.

[196] E. N. Zalta, U. Nodelman, and C. Allen, "Stanford encyclopedia of philosophy," *See http://plato. stanford. edu/. Received September*, 2002.

[197] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's razor," *Information processing letters*, vol. 24, no. 6, pp. 377–380, 1987.

[198] J. Sotos and A. C. of Physicians, *Zebra Cards: An Aid to Obscure Diagnosis*. Mt. Vernon Book Systems, 2006.

[199] 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.

[200] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Machine learning: The high interest credit card of technical debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[201] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature News*, vol. 533, no. 7604, p. 452, 2016.

[202] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725–726, 2018.

[203] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. OConnell et al., "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, p. 203, 2018.

[204] All of Us Research Program Investigators, "The "All of Us" research program," *New England Journal of Medicine*, vol. 381, no. 7, pp. 668–676, 2019.

[205] "The International Parkinson Disease Genomics Consortium (IPDGC)." [Online]. Available: https://pdgenetics.org

[206] 110th Congress of United States of America, "Genetic information nondiscrimination act [hr 493. s. 368]," 2008. [Online]. Available: https://www.congress.gov/bill/110th-congress/house-bill/00493

[207] European Parliament and Council of the European Union, "Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)," 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[208] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[209] "GenoML - Automated Machine Learning (AutoML) for Genomics." [Online]. Available: https://genoml.github.io

[210] "Terra.Bio." [Online]. Available: https://terra.bio

[211] LinkedIn Economic Graph Team, "Linkedin workforce report — united states — august 2018," 2018. [Online]. Available: https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018