MACHINE LEARNING FOR DIGITAL HEALTH: TRANSFORMING CLINICAL DATA INTO KNOWLEDGE

BY

RACHNEET KAUR

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Industrial Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Richard B. Sowers, Chair
Professor Carolyn L. Beck
Professor Thenkurussi Kesavadas
Assistant Professor Manuel E. Hernandez

ABSTRACT

Promoting well being and healthy aging in older adults is becoming increasingly crucial to advance towards a better quality of life. In an effort towards the same, we propose some current and novel approaches to the study of neurological disorders in this thesis. Broadly, we attempt to address recurring problems in neurological disorders, namely early-stage disease diagnosis and progression prediction. Diagnosing neurological conditions is difficult, especially in the early stages, many individuals go undiagnosed partly due to the complex heterogeneity in disease progression. Thus, we study the integration of artificial intelligence (AI) and health data that we believe may provide a viable patient-centric approach to aid clinicians in designing novel AI- based disease prediction strategies and monitoring disease progression. Our ultimate objective is to facilitate the future developments of AI in digital healthcare.

This work proposes new data-driven machine learning-based solutions utilizing health data from multiple modalities, such as gait (e.g., spatiotemporal gait metrics, acceleration, ground re- action forces), cognitive, functional, and longitudinal clinical assessments, and neural responses measured via electrophysiological (e.g., electromyography (EMG), electrocardiography (ECG)) signals, to improve early disease prediction and progression in neurological movement disor- ders. We measure our ability to use these signals to classify disability and predict progression of cognitive and motor changes in persons with neuromuscular disorders. This thesis is a multidisciplinary effort that involves novel combinations of sensors, vision, machine learning, biomechanics, and dynamical analyses to better characterize neurological disorders. These studies on the integration of AI and health data may provide a viable patient-centric approach to aid clinicians in designing novel AI-based disease prediction strategies and monitoring disease progression. This may help providers to individualize treatment plans and design improved clinical trials; thus, help reduce the skyrocketing healthcare costs in the future.

The focus of this dissertation is on the following three areas under the broad umbrella of AI for digital healthcare: 1) Gait analysis for differentiation of neurological disorders, where we focus on disease diagnosis and study gait data-driven methodologies for an automated quantification of neurological gait disorders, such as multiple sclerosis and Parkinson's disease, 2) Clinical data analysis for prediction of disease progression, where we focus on early-stage disease progression

ACKNOWLEDGMENTS

This dissertation is a result of support and encouragements from many incredible people. Firstly, I would like to present my deepest acknowledgment to my advisor Professor Richard B. Sowers, for his continuous backing, encouragement and guidance on my research and many other aspects of my PhD life. Professor Richard is extremely inspirational with great enthusiasm for his re- search and the accomplishment of his research students. This dissertation is only possible with his inspiring research ideas, invaluable insights and very generous time with my research. Pro- fessor Sowers always had a deep belief in me and gave me the freedom to examine and determine new research directions during my time at UIUC.

I am grateful to my mentor and close collaborator Professor Manuel E. Hernandez for his throughout direction and the opportunity of working on very current and impactful health prob- lems. Thank you Prof. Hernandez for always working with me to meet paper deadlines. I also wish to thank other members of my doctoral committee, Prof. Carolyn L. Beck, and Prof. Thenkurussi Kesavadas for their ingenious suggestions on my dissertation. I am grateful for their help, guidance, and support. I also like to thank Prof. Roy H. Campbell, Prof. Justin Sirignano, Prof. Robert W. Motl and Prof. ChengXiang Zhai, for their help and guidance during my re- search and studies at the University of Illinois at Urbana-Champaign. Prof. Sirignano's Deep Learning, Prof. ChengXiang's Information Retrieval, and Prof. Campbell's AI in Healthcare class, have been some of my very favorite courses at UIUC. I would also like to thank Prof. Sirignano for mentoring me on deep learning and high frequency trading research.

I am very happy to have had the opportunity to collaborate with an amazing set of friends and researchers from the Industrial Engineering and Computer Science departments, through my time. I would like to thank all of the undergraduate and graduate students that I worked with, including Zizhang Chen, Joshua Levy, Maxim Korolkov, Rongyi Sun, Yang Hu, Alka Bishnoi, Liran Ziegelman, Faraz Faghri, Vipul Satone and Anant Dadu for their enthusiasm and hard work. I would also like to thank Xiaobo Dong, and Lei Fan for working with me on the high frequency trading research project and multipe semesters of co-teaching the Deep Learning course at UIUC.

I am grateful to have had the William A. Chittenden II graduate fellowship by the Industrial

Engineering department for academic years 2018-2021.

Many thanks to the staff of the Industrial & Enterprise Systems Engineering department at UIUC, including but not limited to: Ann Christine, Lauren Redman, Staci McDannel, Tracey A. Rich, and Holly Kizer, for their prompt help and support at multiple junctures of my studies.

I have had the pleasure to pursue research internships at various companies during my PhD. I would like to thank all my research managers, mentors and colleagues during my summer internships. Specifically, I would like to thank Zhen Zeng at the J.P. Morgan AI Research, Chiranjeet Chetia at the Visa Research, Brian J. Stankiewicz at the 3M AI Research and Matteo Nicoli at the Quantlab Financial, for the enjoyable and technically fulfilling collaboration experiences.

I thoroughly enjoyed my years at UIUC, all thanks to many graduate students and friends. I would especially like to thank Prashant Jayannavar for countless research discussions, life guidance and strong support. Thank you for being my best friend, philosopher and guide. Many thanks to Jaydeep Chanduka and Tarun Giri for many years of friendship and advising during and after UIUC. Thanks for all the constant nagging and effort that you both put in to make me socialize during my first two years at UIUC. I would like to also thank all my "snack group" friends: Vipul Harsh, Amitha Sandur, Atul Sandur, and Shreya Arya, for daily dose of fun and cheerfulness during my PhD journey. Thank you Unnat Jain, Anshika Gupta, Vaishnavi Subramanian, Anand Bhattad, Konik Kothari, Sameer Khan, Anushri Pampari, and many other "core" group members for continuous encouragement and togetherness. Thank you everyone for always reminding me that we are all in this together. I would also like to mention and thank my childhood friend, Manmohit Singh, for providing support and optimism.

But most importantly, this PhD journey was possible and successful only because of the immeasurable support of my family. I dedicate this thesis to my parents, Bhupinder Kaur and Inderjeet Singh, and my brother, Harsimran Singh, for all these years of their unconditional love and support. My mother, Bhupinder Kaur, has been a constant pillar of support in my life and the reason for my strength. Her endless love and care has lead the way for my PhD and life accomplishments. Thank you Harsimran for the tremendous unsaid encouragement, inspiration, and support.

This work was supported in part by the University of Illinois Center for Wearable Intelligent Technologies SRI. This work would not have been possible without the support of the the Illinois Geometry Laboratory of the Department of Mathematics at the University of Illinois. I would also like to thanks JUMP ARCHES for providing financial support for this research. This work utilizes resources supported by the National Science Foundation's Major Research Instrumenta- tion program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. This work also utilizes computational resources supported by the high performance computing center of the MSFE program at the University of Illinois at Urbana-Champaign.

This work was supported in part by the Intramural Research Programs of the National Institute on Aging and the National Institute of Neurological Disorders and Stroke (project numbers: Z01-AG000949-02 and ZIA-NS003154). 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; BioClin- ica, 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 Insti- tutes 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.

I would also like to thank all the participants and their families who collaborated with their time and data for this research.

CHAPTER 1

INTRODUCTION

As 10, 000 Americans are turning 65 each day [1] and nearly one in every five United States residents is projected to be aged 65 and older in 2030 [2], promoting well being and healthy aging in older adults is becoming increasingly crucial to advance towards a better quality of life. In an effort towards the same, we propose some current and novel approaches to the study of neurolog- ical disorders in this thesis. Broadly, we attempt to address recurring problems in neurological disorders, namely early-stage disease diagnosis and progression prediction. Diagnosing neurological conditions is difficult, especially in the early stages, many individuals go undiagnosed partly due to the complex heterogeneity in disease progression. Thus, we study the integration of artificial intelligence (AI) and health data that we believe may provide a viable patient-centric approach to aid clinicians in designing novel AI-based disease prediction strategies and monitoring disease progression. Our ultimate objective is to facilitate the future developments of AI in digital healthcare.

This thesis concentrates on the following three areas under the broad umbrella of AI for dig- ital healthcare. In Part I: Gait analysis for differentiation of neurological disorders, we focus on disease diagnosis and study gait data-driven methodologies for an automated quantification of neurological gait disorders, such as multiple sclerosis and Parkinson's disease. In Part II: Clinical data analysis for prediction of disease progression, we focus on early-stage disease pro- gression prediction and propose machine learning models to identify etiological disease subtypes and study trajectory progression in Alzheimer's disease. Finally, in Part III: Virtual reality for analysing neural responses to anxiety, we examine the potential of virtual reality neurorehabili- tation for ameliorating fall-related anxiety in adults. The remainder of the Introduction chapter is organized as follows. We start out in Section 1.1 by providing some background information and motivation for studying the three above discussed areas. Then, we discuss the overarching aims for this thesis in Section 1.2, followed by some keystone challenges and contributions in Section 1.3. Finally, we outline the organization of this dissertation in Section 1.4.

1.1 Background and Motivation

In this section, we review some background and motivation for the three areas of interest in this thesis, namely Part I: Gait analysis for differentiation of neurological disorders on disease diagnosis in multiple sclerosis and Parkinson's disease (see 1.1.1), Part II: Clinical data analysis for prediction of disease progression on early-stage disease progression prediction in Alzheimer's disease (1.1.2), and Part III: Virtual reality for analysing neural responses to anxiety on the potential of virtual reality neurorehabilitation in human anxiety (1.1.3).

1.1.1 Gait Analysis for Differentiation of Neurological Disorders

In this section, our interest is disease diagnosis and in particular, examining the effectiveness of a gait data-driven framework for multiple sclerosis and Parkinson's disease gait dysfunction prediction.

1.1.1.1 Gaitdata

Gait is a complex dynamical activity which can reflect a range of kinetic, kinematic, and electrophysiological behaviors [3, 4]. A typical walking gait consists of recurrent gait cycles or strides. Figure 1.1 depicts a sagittal plane view of a gait cycle. A single gait cycle consists of the fol-

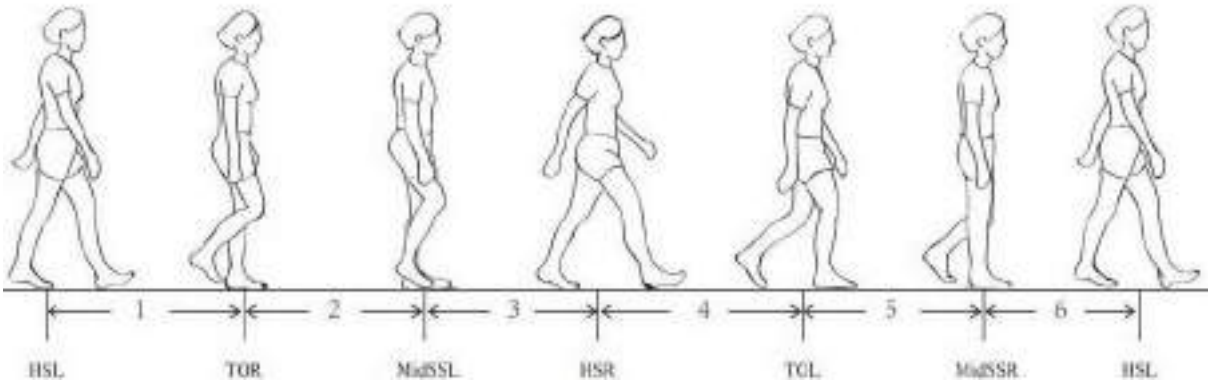

Figure 1.1: Sagittal plane view of a gait cycle

lowing gait events (in order)- HSL: heel strike left, TOR: toe-off right, MidSSL: midstance left, HSR: heel strike right, TOL: toe-off left, MidSSR: midstance right, with the next HSL start- ing a new stride. Both "normal" and "impaired" gait are difficult to model and quantify due to inherent individual variability arising from motor and cognitive differences (cf. [5, 6]). Gait

provides a useful test-bed for understanding how sensing and machine learning can be combined to understand bio-mechanical behavior.

Our efforts will center around detecting and identifying progression of cognitive and motor symptoms in neurological gait disorders, namely, multiple sclerosis and Parkinson's disease. Neurological gait disorders are associated with an increased risk of falls in older adults [7]. Abnormal gait is prevalent in older adults, overall has been observed in 35% of older adults, and associated with a greater risk of institutionalization and mortality [8]. Older adults with chronic neurodegenerative conditions such as Parkinson's disease and multiple sclerosis, com- monly present with gait dysfunction [9, 10], and a high fall risk [11, 12].

1.1.1.2 Multiplesclerosis Affecting approximately 2.8 million people globally [13], multiple sclerosis is a chronic neuro-

logical condition of the central nervous system leading to various physical, mental and psychiatric complexities [14]. Multiple sclerosis presents immensely heterogeneous clinical symptoms among individuals from nothing sometimes to serious signs such as muscle immobility, speech, vision complications, and memory issues in rest [15]. Figure 1.2 depicts a nerve affected by multiple sclerosis. Medical care and additional costs for persons with multiple sclerosis in the

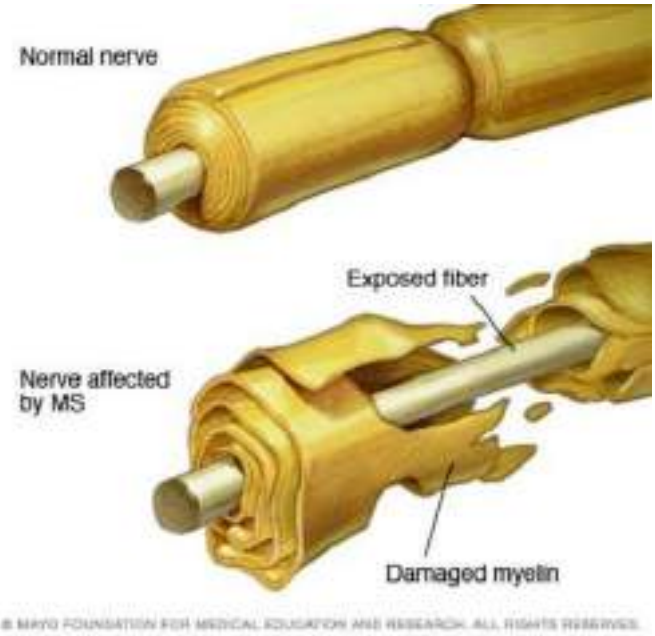

Figure 1.2: Nerve affected by multiple sclerosis

United States alone are around $24 billion annually [16]. With no known cure and clinically unpredictable disease progression, efficient diagnosis of MS in initial stages is vital for clinicians

in defining effective therapy and medication strategies [17].

The occurrence of multiple sclerosis is now highest within 50-60 year old adults [18] and movement limitations are amongst the most frequent and early markers amidst all manifestations [19]. Nearly 85% of persons with multiple sclerosis describe gait disorders as a major complication [9], and roughly 50% patients need walking assistance within 15 years of multiple sclerosis onset [20]. Mobility impairments in persons with multiple sclerosis may further evolve into a fear of falling, significantly impacting physical participation and quality of life [21]. The monitoring of most underlying manifestations of multiple sclerosis require detailed and expensive medical tests in a clinical setting, as opposed to examination of gait impairments, which are economically, quickly and remotely monitorable without the need of a therapist. Real-time gait monitoring, using objective kinematic and kinetic characteristics of gait, may thus be important for identifying people with an increased risk for sudden worsening of the disease. Cognitive impairments in persons with multiple sclerosis are also common [22–24], particularly in processing speed, executive functions, and memory [25–27]. Recent studies [28–34], including the work in mobile neuroimaging [35–37], have confirmed that, similar to aging, attention and executive functions are related to mobility in persons with multiple sclerosis. In particular, shorter strides, decreased cadence, and slower gait speed have been observed in persons with multiple sclerosis when walking and performing concurrent cognitive-taxing tasks [38,39]. Furthermore, cognitive processing speed has been found to be associated with declines in gait performance while walking and talking [38]. Thus, through continuous gait monitoring, both cognitive and motor function changes in persons with multiple sclerosis may be observable.

1.1.1.3 Parkinson'sdisease

In order to characterize gait patterns in neurological population, along with adding persons with multiple sclerosis, we also included persons with Parkinson's disease to our study, i.e., we not only learn distinguishing characteristics between multiple sclerosis and healthy gait but also between neurological gait from multiple sclerosis and another movement disorder; this, in addition to being significantly valuable, is also potentially more challenging.

Parkinson's disease is also a chronic and progressive neurodegenerative disorder of the central nervous system affecting more than 10 million people worldwide [40]; the annual combined direct and indirect costs for persons with Parkinson's disease in the United States are nearly $52 billion [41]. With no known cause or cure and serious disease complications including both motor and non-motor symptoms, investigating additional tools is imperative to aid clinicians in early Parkinson's disease diagnoses.

1.1.2 Clinical Data Analysis for Prediction of Disease Progression

In this section, our interest is disease progression prediction and in particular, examining the ability of machine learning models to identify the unclear etiological disease subtypes in Alzheimer's disease.

1.1.2.1 Alzheimer'sdisease Alzheimer's disease is a common, age-related, neurodegenerative disease that impairs a person's
ability to perform basic activities of daily living. Diagnosing Alzheimer's disease can be challenging, especially in the early stages. Many patients remain undiagnosed, partly due to the complex heterogeneity in disease progression. This diagnostic challenge highlights a need for early prediction of the disease course to assist its treatment and tailor management to the disease progression rate. Recent developments in machine learning techniques provide the potential to predict disease progression and trajectory of Alzheimer's disease and classify the disease into distinct subtypes. In a systematic survey of literature surveying disease subtyping in Alzheimer's disease and related dementias subtyping, we identified nearly fifty independent reports. Most of these reports separately looked at either progression or subtyping but not generally the two together. Additionally many of these past works focused on specific biomarkers, imaging metrics or clinical outcomes in a semi-hypothesis driven manner.

We propose and evaluate a completely hypothesis-free, data-driven effort that incorporates current best practices in longitudinal machine learning (long short-term memory modeling) to predict peri-diagnostic trajectories in Alzheimer's disease with high accuracy in an open science context. This approach may help providers identify distinct disease subtypes with different progression rates and trajectories in the early stages of the disease, allowing for more efficient and personalized healthcare delivery. With additional information about the progression rate of Alzheimer's at hand, providers may further individualize treatment plans. The predictive tests in this study allow for early Alzheimer's disease diagnosis and facilitate the characterization of distinct Alzheimer's subtypes relating to disease progression. These findings are a crucial step forward for early disease detection. These models can be used to design improved clinical trials for Alzheimer's research.

1.1.3 Virtual Reality for Analysing Neural Responses to Anxiety

In this section, our interest is virtual reality-based neurorehabilitation and in particular, it's potential to understand individual cognitive and neural process variations in response to realistic

and complex environments and sensorimotor integration alterations under anxiety-inducing conditions.

1.1.3.1 Fearoffalling

Mobility and in particular fear of falling (FOF) remains one of the significant issues while trying to progress towards healthy aging. Several researches have explored interactions between mobility and healthy aging in the past [42–44]. A study in [45] determined that aerobic training facilitates clinical and physical improvements in depressive elderly adults, and hence protecting against a decrease in cortical activity. Experiments conducted by [46] on 21 elderly subjects aging 67 to 92 years observed that the margins to the spatial-temporal boundaries of postural stability decrease with advancing age, which may contribute to the progressive instability of posture with aging in the elderly. Experiments on 30 older adult subjects in [47] explored difficulty disengaging from fall-threatening stimuli among fall-fearful older adults. FOF, known to reduce balance performance in older adults, is a prominent risk factor for fall risks; which is a leading cause for injuries and mortality among the elderly [48]. It is often accompanied with physical and psychological effects that enhance the fall risk. Given this strong association between the FOF and fall risks [49], description of the mechanisms that govern anxiety-related changes in postural control is critical. Understanding these anxiety-related alterations may lay the foundations for the development of novel therapeutic procedures to help reduce fall risks, especially among the older adults with movement disorders such as Parkinson's disease. As FOF might also alter postural control strategies [50], studying changes in anxiety levels when at different heights and during induced perturbations might further help explain this correlation. Visual stimuli is known to cause anxiety in patients suffering from acrophobia or pathological phobia of heights, while they try to manage standing balance [51]. Acrophobia is one of the most prevalent and severe specific phobias affecting nearly 4.9% of the population [52]. Hence, studying anxiety-related responses to immersive visual stimuli and induced height and perturbation changes will aid in further understanding acrophobic responses. Research on Parkinson's disease patients have deduced that exposure to elevated heights inclines them to possibly fall in a backward or medial-lateral direction [53]. Therefore, understanding responses to height variations is also significant in designing remedial treatments for elderly and patients suffering from Parkinson's disease and other movement disorders [54].

1.1.3.2 Virtualreality

Virtual reality (VR) is an experience where a person interacts with a controlled or modifiable artificial reality environment emulating the real world, where the subject responses can be monitored and evaluated [55]. Further, VR frameworks assist in isolating subjects from non-natural or complex laboratory settings [56,57]. Hence, immersive VR environments, being secure and practicable to provide representative visual scenarios, have been investigated broadly by researchers to examine acrophobia and neural training [58, 59]. A sample VR pathway is shown in Figures 1.3 and 1.4. For people suffering with motor and mental health dysfunctions, VR training is potentially a viable therapy. The underlying idea of VR-based therapeutic treatments for motor and cognitive disorders is to engage subjects in multisensory training, and hence aid to enhance neuroplasticity of the human brain. A comprehensive review of VR as a platform for neuromodulation and neuroimaging was presented by [60]. To demonstrate that VR can induce psychological responses analogous to real world height controls for studying fall-related anxiety in older adults, a system that records neural, physiological, and behavioral data in an engaging virtual environment, while implementing sensory and mechanical perturbations corresponding to high postural threat conditions in the real world, was introduced in [61]. A realistic ocular feedback of cliff scenarios in VR has been studied for analyzing acrophobic responses of the human brain [62]. To understand dynamic neurological responses to visual cliffs while walking for mitigating FOF, especially among the elder population and those suffering with movement disorders, an experimental setup that monitors a subject's real time neural activity via electroencephalogram (EEG) signals while being immersed in a virtual world and walking on an instrumented treadmill, was investigated by [63]. Results in [64] on 10 healthy young adults indicated an influence of psychological factors related to postural threat on the cortical activity associated with postural reactions to unpredictable perturbations. Further, a study conducted by [65] on 10 healthy individuals found insights into links between cortical and cognitive influences on compensatory balance control. Fall risk and postural stability experiments in [66] concluded the efficacy of a short-term VR-based balance training program on the balance ability of patients suffering with multiple sclerosis. The study concluded that the fall risk index and overall stability index of the patients significantly improved after 24 sessions of the VR balance training.

1.1.3.3 Braincomputerinterfaces

To investigate anxiety responses and neurological feedback, some researchers have previously examined experimental setups unifying VR and EEG. Consequences of high heights exposure

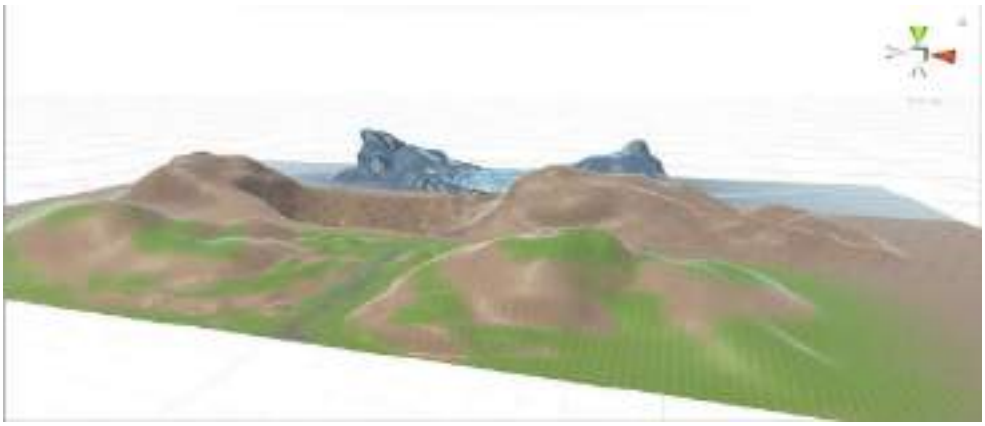

Figure 1.3: Sample Terrain: Overview

Figure 1.4: Sample Terrain: Subject view

during beam-walking in a VR environment on physiological stress and cognitive loading was explored using statistical analysis on EEG signals in [67]. In experiments conducted by [68], a combination of VR with robotic-based rehabilitation induced an improvement in gait and balance among patients suffering with chronic hemiparesis. Statistical analysis on the EEG data collected during the experiments suggested that the use of VR may entrain brain areas responsible for motor planning and learning, and hence may potentially lead to an enhanced motor performance in humans. Moreover, EEG and VR-based brain-computer interface (BCI) systems, that allow brain responses to control virtual robots or surroundings, may serve in further facilitating neural rehabilitation. Figure 1.5 illustrates the BCI setup for an example VR walking experiment. The work proposed in [69] concluded that a closed-loop EEG-based BCI-VR system enhances cortical involvement in human treadmill walking, triggers cortical networks involved in motor learning, and further enhances voluntary control of human gait. This study indicates that EEG has the ca- pacity to monitor cortical activity in treadmill walking, hence enabling EEG-based BCI systems for walking as a paradigm for improving the rehabilitation adequacy. A few other VR inspired

walking based BCI systems have been examined in the past for neurorehabilitation [69–71].

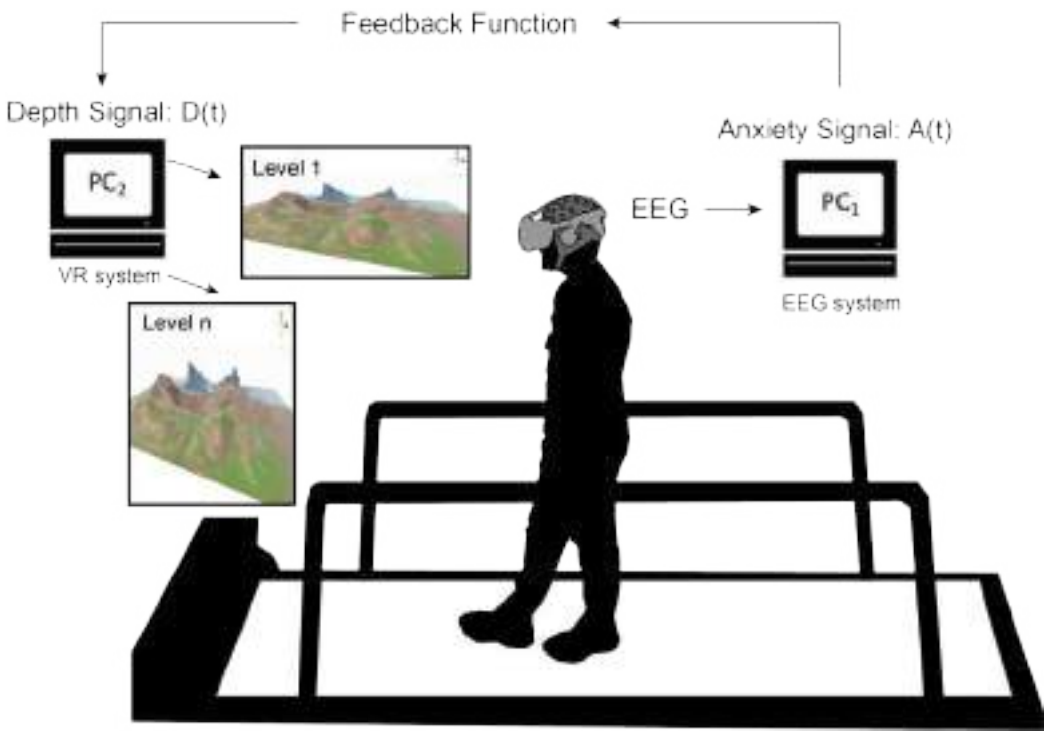

Figure 1.5: Brain-computer interface setup for an example VR walking experiment. All subjects walk at a comfortable pace on a treadmill adaptive to their speed, while being immersed in a virtual world. The cliffs in the terrain modify depths to increase or decrease the scariness in response to the subject's real time neural responses.

1.2 Objectives

Diagnosing neurological disorders is difficult, especially in the early stages; many individuals go undiagnosed partly due to the complex heterogeneity in disease progression. This work proposes new data-driven machine learning-based solutions utilizing health data from multiple modalities, such as gait, cognitive, functional, and longitudinal clinical assessments, and neural responses measured via EEG, to improve early disease prediction and progression in neurological movement disorders. Further, we discuss the challenges and considerations that need to be addressed for the future of AI in digital healthcare. The overarching aims for this thesis are follows:

- Understand how machine learning can be applied to dynamic health data

- Facilitateearlydiseasepredictionanddiseaseprogressionpredictioninneurologicalmovement disorders
- Propose open source, interpretable, accurate, and rapid remote monitoring solutions to automate disease prediction and progression prediction assessments

These studies on the integration of AI and health data may provide a viable patient-centric approach to aid clinicians in designing novel AI-based disease prediction strategies and monitoring disease progression. This may help providers to individualize treatment plans and design improved clinical trials; thus, help reduce the skyrocketing healthcare costs in the future. The data collection for these works were done at the Mobility and Fall Prevention Research Laboratory in the Department of Kinesiology & Community Health within the College of Applied Health Sciences at the University of Illinois at Urbana-Champaign. In an attempt to support open science, we made code base for all works discussed in this thesis publicly available so that other researchers can improve upon it over time as larger patient cohorts become available.

1.3 Keystone Challenges and Contributions

In this thesis, we addressed multiple technical and clinical challenges. Below, we highlight some of these major challenges and our approach towards addressing them.

- Small sample sizes, and particularly with the unhealthy population group. We attempt to extract statistical insights from admittedly a small number of test subjects in order to mitigate some of the concerns related to small sample sizes.

- Multi-modal datasets. Most current applications of AI in medicine utilize only one data modality as opposed to clinicians that analyse multiple data modalities for disease diagnosis and evaluation. We seek to integrate several data modalities, such as clinical, kinetic, kinematic, electrophysiological, and imaging data, in an attempt to develop multi-modal solutions that might capture the inherent heterogeneity and complexity present in the neurological disorders.

- Pre-existing useful domain knowledge. We have performed domain-knowledge based feature engineering to take advantage of the useful pre-existing specialized discipline information that might be present in the collected health data.

- Explainability of machine learning and deep learning models. It is sometimes difficult to interpret the reasoning behind the outcome generated by the “black box” AI models. In

this work, we have thoroughly explored the interpretability of our optimal models via post hoc analysis to generate model explanations. This is especially useful to establish trust in our predictions from complex, but more accurate deep learning models.

• Absence of large amounts of labelled data for supervised learning. We did not have large and reliable amounts of labelled data to design precise subtype prediction models for neurological population using supervised learning approaches. To address this, we used unsupervised clustering techniques, such as Gaussian mixture models, to identify and label subtypes of the disease.

• Data Dimensionality. With extracted features from the multiple cognitive, functional and longitudinal data assessments, we had a multi-dimensional space that captured both the features of the disease and the progression rate of these features. Thus, to create meaningful low-dimensional representations, we used data dimensionality reduction methods, such as non-negative matrix factorization, that summarize these extensive data measures.

• Evaluation of models beyond accuracy. Since using only accuracy as the evaluation metric for our machine learning models may not be enough, we have used sensitivity and specificity to better gauge the performance of our work in a clinical setting.

The main contributions of this thesis are the following.

- Propose a spatiotemporal and kinetic gait features-based machine learning framework for multiple sclerosis prediction
- Improveupontheperformanceofthisdomainknowledge-basedframeworkusingdynamics from multiple strides and deep learning
- Propose a multi-view gait video data-driven deep learning methodology for multiple sclerosis prediction
- Study the feasibility of wearable sensors to identify older adults with balance dysfunction
- Identify disease onset and trajectory progression in Alzheimer's disease utilizing clinical data and data-driven models
- Extract information from physiological responses to virtual reality-based visual stimuli in order to reflect low or high anxiety inducing states

1.4 Dissertation Outline

This thesis is organized as follows.

Part I: Gait analysis for differentiation of neurological disorders consists of Chapters 2-5, that propose and evaluate a gait data-driven methodology for an automated and objective quantification of neurological conditions, such as multiple sclerosis and Parkinson's disease.

• Chapter 2 proposes a spatiotemporal and kinetic gait features-based machine learning framework for multiple sclerosis prediction [72]. We evaluate the effectiveness of the proposed methodology so as to generalize across different walking tasks and subjects after gait normalization. We also study the explainability of our machine learning models via post-hoc analysis and application of gait features in learning progression space in persons with multiple sclerosis.

• Chapter 3 proposes a multi-view gait video data-driven deep learning methodology that classifies strides of persons with multiple sclerosis, healthy older adults and persons with Parkinson's disease [73]. We evaluate the effectiveness of the same to generalize across different walking tasks, subjects, and both together. The studied workflow is convenient, low-cost, accurate, serves as a rapid remote monitoring tool, and is contact-less to provide convenience and automaticity in gait assessments in the wild.

• Chapter 4 analyses the additional information from gait dynamics and variations across temporally ordered strides, and proposes a deep learning-based methodology to classify multi stride sequences of persons with multiple sclerosis from healthy older adults [74]. This framework improves upon the performance of domain knowledge-based framework in Chapter 2 across both task and subject generalization model designs.

• Chapter 5 studies and validates the feasibility of wearable sensors to identify older adults with balance dysfunction [75]. We use accelerometer data from hip and knee motion during walking to analyse high or low dynamic balance ability, as labelled by a motor control test.

Part II: Clinical data analysis for prediction of disease progression consists of Chapter 6, that studies disease onset and progression in Alzheimer's disease.

• Chapter 6 proposes a completely hypothesis-free, data-driven effort that incorporates current best practices in longitudinal machine learning (long short-term memory modeling) to predict peri-diagnostic trajectories in Alzheimer's disease with high accuracy in an open science context [76,77]. We identify and accurately predict three subtypes of Alzheimer's disease with varying rates of disease progression. In addition, we provide an interactive website, i.e. for clinical researchers to predict the clinical subtype of an Alzheimer's disease patient based on clinical parameters. We have identified a fast progressing subset of the Alzheimer's population that may be optimal for inclusion of clinical trials. Faster progressing cases may likely be more ideal for enrollment as they can potentially show successful drug readouts in a shorter period of time thus allowing for shorter, more ef-

Part III: Virtual reality for analysing neural responses to anxiety consists of Chapters 7 and 8, that examine the potential for virtual reality neurorehabilitation aimed at ameliorating fall-related anxiety in adults.

• Chapter 7 studies physiological responses to virtual reality-based visual stimuli to re- flect low or high anxiety-inducing states [63, 78–80]. This is useful for applications in human postural control; and is a step towards a self monitoring and regulated person- alized virtual reality environment that allows subjects to learn to decrease their anxiety in balance demanding conditions.

• Chapter 8 proposes a deep learning classifier to automatically identify brain components in the independent components extracted from the subject's electroencephalography (EEG) data gathered while they are being immersed in a virtual reality environment [81]. EEG analysis demands substantial training and time for removal of distinct unwanted independent components, generated via independent component analysis, corresponding to artifacts. The considerable subject-wise variations across these components motivates defining a procedural way to identify and eliminate these artifacts.

Part IV: Conclusion consists of Chapter 9, that summarizes our overall contributions and concluding remarks.

• Finally, Chapter 9 highlights concluding remarks along with some potential future directions for work proposed in this thesis.

CHAPTER 2

LEARNING A DOMAIN KNOWLEDGE-BASED FRAMEWORK

In this chapter, we review the work in Predicting Multiple Sclerosis from Gait Dynamics Using an Instrumented Treadmill - A Machine Learning Approach.

2.1 Introduction

Multiple Sclerosis (MS) is a chronic demyelinating and neurodegenerative disorder that impairs the central nervous system. It can affect a range of cognitive, physical, and psychiatric processes [14, 83]. Severe symptoms include impairment of vision and sensory abilities, muscle paralysis, and depression [15], with mobility impairments being one of the most frequent signs [19]. MS affects approximately 1 million people in the United States and more than 2.5 million globally [84]. Peak prevalence is in adults 50-60 years of age [18]. Direct medical treatment expenses and indirect costs in terms of lost productivity, additional need for caretakers, and amenities for persons with MS (PwMS) are estimated to be $24 billion annually in the United States [16].

Walking and balance difficulties are some of the most common indicators in PwMS; nearly 85% of PwMS describe gait disorders as a major complication [9] and roughly 50% patients need walking assistance within 15 years of MS onset [20]. Secondary effects often include fear of falling, significantly impacting the quality of life of PwMS [21]. In contrast to monitoring of most underlying manifestations of MS, which require neurological examinations by a trained practitioner, gait can be quickly and remotely monitored. Thus, objective gait monitoring, which expands upon typical clinical tests [85], may be important for designing disease prediction and progression strategies in PwMS. Past research on MS assessment with gait-related dynamics has typically relied upon statistical inferences that may not be able to accommodate the heterogeneity present in the disease [86–90]. Given that subtle and heterogeneous patterns of gait changes may arise in PwMS over time, a machine learning (ML) approach will be valuable for monitoring MS-related changes in older adults.

This study aims to examine MS and disability related changes in spatiotemporal and kinetic gait features after normalization; and evaluate the effectiveness of a gait data-based machine learning (ML) framework for MS prediction (GML4MS), an ML-based methodology to classify strides from older PwMS and healthy controls, so as to generalize across different walking tasks and subjects after gait normalization. Building upon prior work examining MS-related variations in gait characteristics [91], we categorized PwMS using the following two classification designs (see Figure 2.1):

1) Taskgeneralizationestablishingthegeneralityoverdifferenttasks. Wetrainbinary(healthy vs. MS) classifiers on walking (W) trials (tasks) and apply them to walking-while-talking (WT) trials. Task generalization results will hopefully reflect how classifiers trained in supervised lab conditions might work in real-world gait tasks with challenges of divided attention. To monitor disease progression and relapses, task generality is vital as normative data collected in a clinic or lab could be used as a basis to assess gait data collected using wearable sensors in a home-based setting.

2) Subject generalization demonstrating the generality over different subjects. We train binary (healthy vs. MS) classifiers (with a balanced collection of W and WT tasks) on some test subjects and apply them to a withheld separate set of test subjects. These results may have implications in detection of MS in new patients.

2.1.1 Related Work

Several studies have identified declines in gait performance in PwMS, particularly as disabil- ity increases [86, 92, 93]. The increased gait variability and timed 25-foot walk, an objective measure of straight-line gait speed, are standard measures considered for research and clinical gait assessment in PwMS [85]. Most gait-based methods for identifying MS have relied upon traditional statistical techniques such as t-test, and ANOVA (analysis of variance), to examine differences in spatiotemporal features and correlations with disability [86–90]. Supervised ma- chine learning methodologies such as random forest and artificial neural networks have already been used in human gait analysis across other neurological populations [94, 95]. A few prior works have explored machine learning to classify MS using gait data [96], [97]. However, to the best of our knowledge, there is no study utilizing machine learning on spatiotemporal and kinetic gait characteristics for MS prediction. Despite model-based statistical practices present- ing transparency and explainability regarding the contribution of independent features, machine learning approaches may improve performance by addressing high-dimensional and non-linear

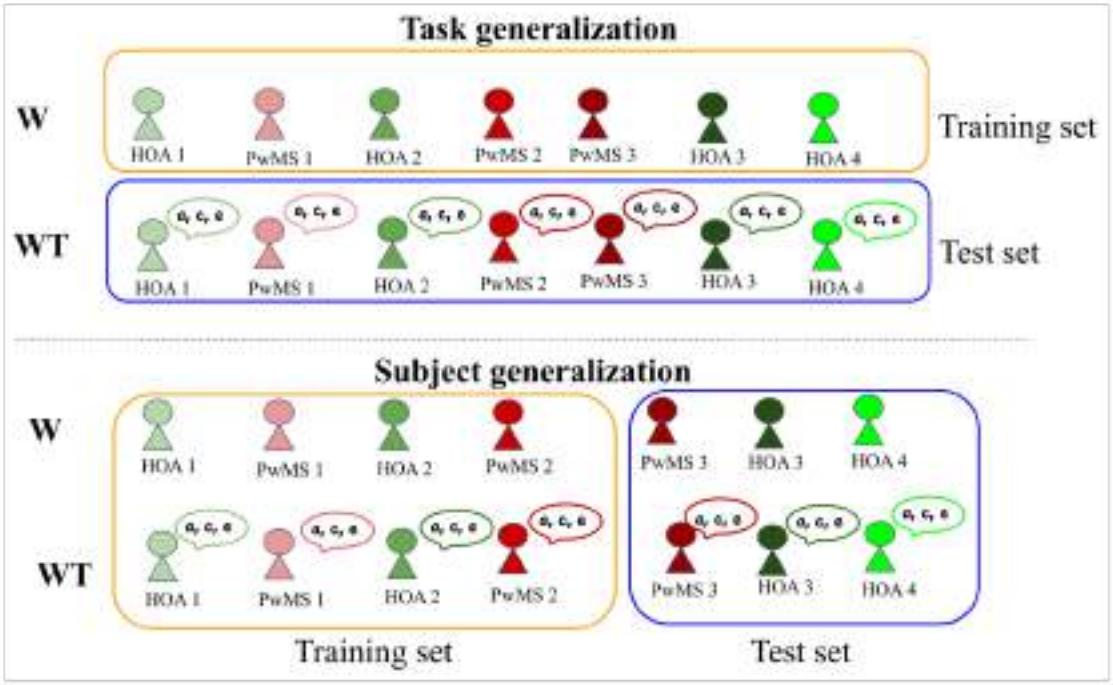

Figure 2.1: Top: Task generalization model, Bottom: Subject generalization design. Healthy older adults (HOA) and PwMS are depicted in shades of green and red, respectively. The indices (1, 2, 3, . . .) along with HOA and PwMS are used as a reference for dummy subjects identifiers.

feature interactions in a model-free way. Further, transforming statistical inference to prediction classes requires defining sensitive classification thresholds.

Distinctive physical characteristics across subjects inherently increase the variability in raw gait parameters and thus limit the efficiency of learning reliable trends in gait features that differentiate healthy and pathological gait [98]. Referring to the performance improvements in previous studies examining neurological diagnosis [95, 99], two normalization strategies, namely size-N (standard body size-based normalization) and regress-N (regression-based normalization using scaling factors derived by regressing gait features on multiple subject demographics) were explored to minimize the dependency of gait features on the subject demographics.

The proposed application of machine learning classifiers to recognize gait patterns of PwMS across tasks and over new subjects is a step forward towards the identification of a tipping point for older people with MS, and worsening of symptoms in the near term. Moreover, we discuss the importance of spatiotemporal and kinetic features, encompassing valuable domain knowledge, in the classification performance. Attributing to prior evidence of gait changes with MS impairment [86, 100, 101], we construct an MS progression space by unsupervised clustering of reduced gait feature space for PwMS to examine the relative correspondence of the defined subgroups to disease severity. This analysis may facilitate strategies to monitor disease progression and

evaluate the effectiveness of disease modifying interventions. The proposed methodology is an advancement towards developing an assessment marker for medical professionals to early diagnose older PwMS who are likely to have a worsening of symptoms in the near term.

The remainder of the chapter is organized as follows. In Section 2.2.1, we introduce the data acquisition paradigm and study participants. In Section 2.2.2, we discuss the data analysis methodology, including feature extraction, data normalization, statistical investigations, ML- based classification and evaluation strategies. In Section 2.3, we present our major statistical findings and prediction model results. In Section 2.4, we examine feature importance, progres- sion space in MS along with some limitations and future directions of our work. Finally in Section 2.5, we highlight the concluding remarks.

2.2 Methods

2.2.1 Experimental Design: Setup and Subjects

The protocol for this study was approved under the University of Illinois at Urbana-Champaign Institutional Review Board number 15674 on 4/3/2015.

2.2.1.1 Experimentalparadigm An instrumented treadmill (C-Mill, Motekforce Link, Culemborg, The Netherlands) in self-
paced mode was utilized to allow subjects to walk at their natural speed. To allow for unbiased force recordings, subjects were instructed to restrain from holding the handrails while walking on the treadmill. For safety purposes, all subjects wore a ceiling-mounted harness and had access to an emergency stop button during all the walking trials. Figure A.1 in the Appendix illustrates the gait data acquisition setup. All subjects walked one trial under two different task conditions, namely single-task condition, W and dual-task paradigm, WT. For the WT task, subjects were asked to walk while reciting alternate letters of the alphabet (i.e. a, c, e, ...), coordinating equal attention between mobility and the cognitive interference exercise to depreciate the influence of task prioritization. The divided attention dual-task walking in a laboratory environment has been demonstrated to be more analogous (as compared to usual walking) to every-day walking in the older adults and hence provides a competent framework to generalize adequacy towards daily-living gait for 24/7 monitoring scenarios [102]. Further, the attention demanding WT task has been examined by researchers for practical implications in designing mobility risk assessment

procedures and predicting the risk of falls and fall-related injuries in older adults and individu- als with other cognitive or movement disorders [103]. For each trial, subjects were instructed to walk at a comfortable pace for up to 75 seconds, after being provided with a brief training session. CueFors 2 software [104] was used to collect gait event data (i.e., left and right heel strike, mid-stance, and toe-off position coordinates and time stamps) and raw data (i.e., vertical ground reaction forces, treadmill speed and center of pressure (CoP) position coordinates at a 500 Hz frequency) during each walking trial. To facilitate the online identification of gait events, an online pattern recognition algorithm detects maxima and minima in the butterfly patterns (see Section 2.2.2.2) of the center of pressure profiles, that are collected in real time via a force plate embedded in the treadmill [105]. Table A.1 in the Appendix describes the collected raw features.

2.2.1.2 Studyparticipants

Twenty individuals from each cohort, MS patients (age: 61.05

weight: 74.89

and healthy older adults (HOA) (age: 61.2

± 1.62 [1.0 – 6.0] as evaluated by the Kurtzke's Expanded Disability Status Scale (EDSS) [107]), were relapse-free for at least a month prior to experimental trials, and had no other cognitive dysfunction or neurological disorders. EDSS, monitoring sensory, motor, brain stem, visual, cerebellar, bowel and bladder, pyramidal and other functions, is an accepted method to quantify disability in PwMS. For this work, we divided PwMS into three sub groups based on their EDSS score: mild (1.0-2.5), mild-to-moderate (3.0-4.5) and moderate (5.0-6.0). No significant differences (at significance level $\alpha = 0.05$) in age, weight, height, gender and education levels were observed between the two cohorts. Two HOA and three PwMS were excluded from the analysis for holding the handrails (biasing the raw force data).

2.2.2 Data Analysis

2.2.2.1 Gaitterminologyandmathematicalnotation A typical walking gait consists of recurrent gait cycles. A gait cycle or stride is measured from
a foot's heel strike to the subsequent heel strike of the same foot. For our analysis, a stride was characterized by the following gait events: HSR: heel strike right, TOL: toe-off left, MidSSR: midstance right, HSL: heel strike left, TOR: toe-off right, MidSSL: midstance left, with the next HSR starting a new stride. A stride is a consolidation of two steps (i.e. HSR-HSL and HSL-HSR), where a step is marked from a foot's heel strike to the following heel strike of the opposite foot. Figure A.2 demonstrates the longitudinal plane view of a gait cycle. The following are frequently used mathematical notations:

- Let Ns be the total number of valid strides recorded during a subject's complete walking trial on the treadmill
- Let(S, ≤s)where def
  the complete walk where sm

  sm≤s sn if m ≤ n defines a natural order on S. Clearly, cardinality | S | = Ns.
- Let (E, ◁) be an ordered set of six gait events observed during a stride

$$E \overset{def}{=} \{HSR, TOL, MidSSR, HSL, TOR, MidSSL\}$$

  where the order ◁ is defined as follows:

$$HSR \triangleleft TOL \triangleleft MidSSR \triangleleft HSL \triangleleft TOR \triangleleft MidSSL$$

- Let $T_{raw} \overset{def}{=} \{\delta t, t=0,1,2,\ldots,T_{walk}\}$
  and center of pressure recordings where δ = 0.002, Twalk = 75 since the raw data is collected every 0.002 s and each trial lasted for 75 s. For each time stamp $t \in T_{raw}$, define:
  - $S(t)$ asthetreadmillspeed(inm/s)
  - $F_z(t)$ asthegroundreactionforce(inNewton(N))
  - $(CoPX(t), CoPY(t))$ asthecenterofpressurepositionsinxandy-directions(in m)

• Define the Cartesian product $(E \times S, \prec)$ where $E \times S = \{(e,s): e \in E \text{ and } sk \in S\}$ as the set of ordered pairs (e, sk) corresponding to event e of stride sk for every $e \in E$ and $sk \in S$ where eq. 2.1 defines the ordering on $E \times S$.

$$(e, sm) \prec (f, sn) \text{ if } \begin{cases} sm < sn & \text{for } m=n \\ e \lhd f & \text{for } m=n \end{cases} \quad (2.1)$$

For each gait event and stride $(e, sk) \in E \times S$, define:

– (sTk) as the elapsed time (in s) from the start of data recording to (e,sk)

– $(Xe^{(sk)}, Ye^{(sk)})$ as the x and y-coordinates (relative to origin of the treadmill) for the detected (e, sk)

– $Te^{(sk)} \overset{def}{=} \min\{t: t > Te^{(sk)} \text{ and } t \in Traw\}$ as the closest time in Traw (corresponding to the recorded raw forces and center of pressure positions) to the marked time $Te^{(sk)}$

– $FZ^{(e,sk)} \overset{def}{=} FZ(Te^{(sk)})$ as the reaction force at (e,sk)

– $\overline{CoP}_{e,f}^{(sk),(sm)}$ as the center of pressure trajectory between (e,sk) and $(f,sm) \in E \times S$ (events e and f of strides sk and ms respectively) where $(e,sk) \prec (f,sm)$ (eq. 2.1)

$$\overline{CoP}_{e,f}^{(sk),(sm)} \overset{def}{=} \{(CoPX(t), CoPY(t)): Te^{(sk)} \le t \le T_f^{(sm)}\}$$

2.2.2.2 Gait feature engineering for MS characterization To examine cohort related variations in the gait patterns, characteristic kinematic and kinetic features were extracted across strides from the raw gait data using Python 3.6 (see Figure A.3 in the Appendix for our workflow pipeline). The derived features can be categorized as follows:

• Temporal features: 7 temporal gait features, namely stride time, stance time, swing time, supporting (right single, initial double and terminal double) times (in s) and cadence (in steps/min) were computed for each stride. See Figure A.2.

– Stride time is the time between two successive heel strikes of the same foot i.e. HSR-HSR. ST (eq. 2.2) denotes the set of stride times for a complete trial.

$$ST = \{T(sk): sk \in S\} \text{ where } ST(sk) = T_{HSR}^{(sk+1)} - (sTk) \quad (2.2)$$

– Stance time $(StT^{(sk)} = T_{TOR}^{(sk)} - T_{HSR}^{(sk)})$ is the time between heel strike and toe-off (from stride $sk \in S$) of the same foot i.e. HSR-TOR.

– Swing time ($wT^{(sk)} = THSR^{(sk+1)} - TTOR^{(sk)}$) is measured between the toe-off (TOR, sk) and heel strike (HSR, sk+1) of the same foot.

– Support can be categorized as single or double depending on whether only one or both of the subject's feet are in contact with the treadmill's belt, respectively. Single support can further be classified as left/right depending on which one foot supports the subject's body.

*Left single supporting time ($SSL^{(sk)} = THSR^{(sk+1)} - TTOR^{(sk)}$) is the time between toe-off (TOR, sk) and heel strike (HSR, sk+1) of the right foot for stride sk $\in S$.

This is identical to swing time.

*Right single supporting time ($SSR^{(sk)} = THSL^{(sk)} - TTOL^{(sk)}$) is the time between toe-off (TOL, sk) and heel strike (HSL, sk) of the left foot for stride sk $\in S$.

Double support can be identified as initial/terminal based on its onset in the stance phase.

*Initial double supporting time ($DSI^{(sk)} = TTOL^{(sk)} - THSR^{(sk)}$) is the time between heel strike of supporting foot and toe-off of other foot i.e. HSR-TOL from stride sk $\in S$.

*Terminal double supporting time ($DST^{(sk)}$) is calculated between heel strike of the other foot and toe-off of the supporting foot i.e. HSL-TOR from stride sk.

$$DST = DST^{(sk)} : sk \in S \text{ where } DST^{(sk)} = TTOR^{(sk)} - THSL^{(sk)}$$

– Cadence ($C^{(sk)} = 60 \times 2/(THSR^{(sk+1)} - THSR^{(sk)})$) is the walking rate or number of steps taken in a minute (min) i.e. twice the inverse of stride time (in min) for stride sk $\in S$.

• Spatial features: The stride-wise extracted 4 spatial (distance dimension) gait attributes included stride length, stride width (in m) and the dimensionless left and right foot progression angles. See Figure A.4 in the Appendix. Since the foot comes back to its initial position after each stride while walking on a treadmill belt, the y-coordinate of position for the current and next stride event, HSR for instance, will be approximately the same each time. Therefore, to report accurate spatial measures, y-position coordinates were corrected to ac

$BT((e,s_m),(f,sn)) = \int_{t1=Te}^{t2=Tf} (sm)S(t)dt$ is computed as the area under the speed-time curve bounded by the closest times (corresponding to recorded speeds) to the marked times of gait events (e, sm) and (f, sn) $\in E \times S$ where (e, sm) $\prec$ (f, sn) and dt = 0.002. The above integral is numerically approximated via the trapezoidal rule. Hence, the relative

count for the relative treadmill belt travel (BT). Mathematically,

y-coordinate for (f, sn) w.r.t (e, sm) is given by eq. 2.3.

$$\overline{Y}_f^{(sn)} = Y_f^{(sn)} + BT((e,sm),(f,sn)) \quad (2.3)$$

Now, let's define the derived spatial gait markers.

– Stride length (SL(sk)) is the horizontal distance in the walking plane between two subsequent heel strikes of the same foot i.e. between (HSR, sk) and (HSR, sk+1).

$$SL = \{SL(sk) : sk \in S\} \quad \text{where } SL^{(sk)} = \overline{Y}_{HSR}^{(sk+1)} - Y_{HSR}^{(sk)}$$

where $\overline{Y}$ are adjusted for belt travel relative to (HSR, sk).

– Stride width (SW(sk)) is the medio-lateral distance between the two feet i.e. the perpendicular distance between the line connecting two consecutive heel strikes of the same foot i.e. (HSR, sk) and (HSR, sk+1) and the heel strike of the contralateral foot i.e. (HSL, k)s

$$SW(sk) = \frac{1}{D^{(sk)}} | (X_{HSR}^{(sk+1)} - X_{HSR}^{(sk)})(Y_{HSR}^{(sk)} - \overline{Y}_{HSL}^{(sk)}) - (X_{HSR}^{(sk)} - X_{HSL}^{(sk)})(\overline{Y}_{HSR}^{(sk+1)} - Y_{HSR}^{(sk)}) |$$

where $D^{(sk)} = \sqrt{(X_{HSR}^{(sk+1)} - X_{HSR}^{(sk)})^2 + (\overline{Y}_{HSR}^{(sk+1)} - Y_{HSR}^{(sk)})^2}$ and $\overline{Y}$ are adjusted for belt travel relative to (HSR, sk).

– Foot progression angle (FPA) for the right/left foot (sk) (sk) between the progression vector (PR/PL) (joining two consecutive heel strikes of the right/left foot) and the foot vector (FR/FL) (drawn between the right/left foot's heel strike and toe-off) for stride sk [108]. Since staggered walking in PwMS might show significant fluctuations in FPAs, we selected it as a potential feature correlating to MS gait. Mathematically, we have:

$$\theta_* = \{ \theta_*^{(sk)} = (-1)^x \tan^{-1}\left(\frac{YP_*^{(sk)}}{XP_*^{(sk)}}\right) + \ldots (-1)^y \tan^{-1}\left(\frac{YF_*^{(sk)}}{XF_*^{(sk)}}\right) : sk \in S \}$$

$$(XF_*^{(sk)}, YF_*^{(sk)}) = (X_{TO*}^{(sk)} - X_{HS*}^{(sk)}, \overline{Y}_{TO*}^{(sk)} - Y_{HS*}^{(sk)}) \text{ and}$$
$$(XP_*^{(sk)}, YP_*^{(sk)}) = (X_{HS*}^{(sk)} - X_{HS*}^{(sk-1)}, \overline{Y}_{HS*}^{(sk)} - Y_{HS*}^{(sk-1)})$$

$* -$

where ∇ indicates adjusted y-coordinates relative to the belt travel (eq. 2.3), x^* indicates left (L) or right (R) and sk1 denotes sk+1 and sk for L and R, respectively. Exponents x, y are defined as 1, 2 respectively for L and 2, 1 respectively for R. Figure A.5 summarizes these definitions on an overground view of the gait cycles.

• Spatiotemporal features: Derived from the above defined temporal and spatial features, 2 additional spatiotemporal markers, namely stride speed (in m/s) and walk ratio (in m/strides/min) were defined for each gait cycle.

– Stride speed (SS(sk) = SL(sk)/ST(sk)) is defined as the ratio of stride length and stride time for strides $sk \in S$.

– Walk ratio (W(sk) = 2 $\times$SL(sk)/C(sk)) is computed as the ratio of stride length to the number of strides walked per minute (i.e. half the cadence) for gait cycles $sk \in S$.

• Kinetic features: 8 kinetic gait parameters, namely the six forces, one at each gait event (in N) and two butterfly diagram-based features (in m) were identified for each gait cycle.

– Forces $(F_Z^{(e,sk)})$ at each of the six gait events ($e \in E$) were recorded for every stride sk. Thus, for a trial, we have

$$F_{eZ} = \{F_Z^{(e,sk)} : (e,sk) \in \{e\} \times S\} \; \forall \; e \in E$$

– A butterfly diagram reflects the repeated center of pressure trajectory for multiple continuous strides during a subject's walk. The butterfly diagram derived features, especially in the anterior-posterior and lateral directions, have been associated with important neurological functions in PwMS [109] (Figure 2.2). First, the intersection point of the center of pressure trajectory for stride $sk \in S$ is calculated as $(CoPX_{ip}^{(sk)}, CoPY_{ip}^{(sk)})$. Then, the lateral and anterior-posterior shift in the intersection point for a trial are given by

$$\beta_L = \{\beta_L^{(sk)} : s_k \in S\}, \beta_{AP} = \{\beta_{AP}^{(sk)} : s_k \in S\}$$

Define $(CoPX_{ip}, CoPY_{ip}) = \left(\frac{\sum_{k=1}^{Ns} CoPX_{ip}^{(sk)}}{Ns}, \frac{\sum_{k=1}^{Ns} CoPY_{ip}^{(sk)}}{Ns}\right)$ as the mean intersection point. The set of lateral and anterior-posterior squared deviation from the mean intersection point for a trial are given by

$$\alpha_L = \{\alpha_L^{(sk)} = (CoPX_{ip}^{(sk)} - CoPX_{ip})^2 : s_k \in S\}$$
$$\alpha_{AP} = \{\alpha_{AP}^{(sk)} = (CoPY_{ip}^{(sk)} - CoPY_{ip})^2 : s_k \in S\}$$

The lateral (ηL) and anterior-posterior (ηAP) asymmetry can then be defined as the mean lateral and anterior-posterior shift in the intersection points, respectively. Similarly, the lateral (σL) and anterior-posterior (σAP) variability are defined as the lateral and anterior-posterior standard deviation in the intersection points, respectively. We selected βL and αL as the two characteristic features of machine learning variability for our analysis. The significant overlap between the averaged gait cycle waveforms of healthy and MS groups in Figure A.6 depicts the challenges while classifying between healthy and MS gait.

Note that all features except the FPAs are always non-negative. Before deriving the stride-wise features, gait cycles with missing or invalid gait events were eliminated. Since several features, namely stride, swing times, stride length, width and angles generate erroneous estimates for nonconsecutive strides, such values were dropped during data processing. Overall, 1654 (HOA: 905, PwMS: 749) and 1576 (HOA: 878, PwMS: 698) strides are retrieved from W and WT trials, respectively, across 35 subjects (HOA: 18, PwMS: 17).

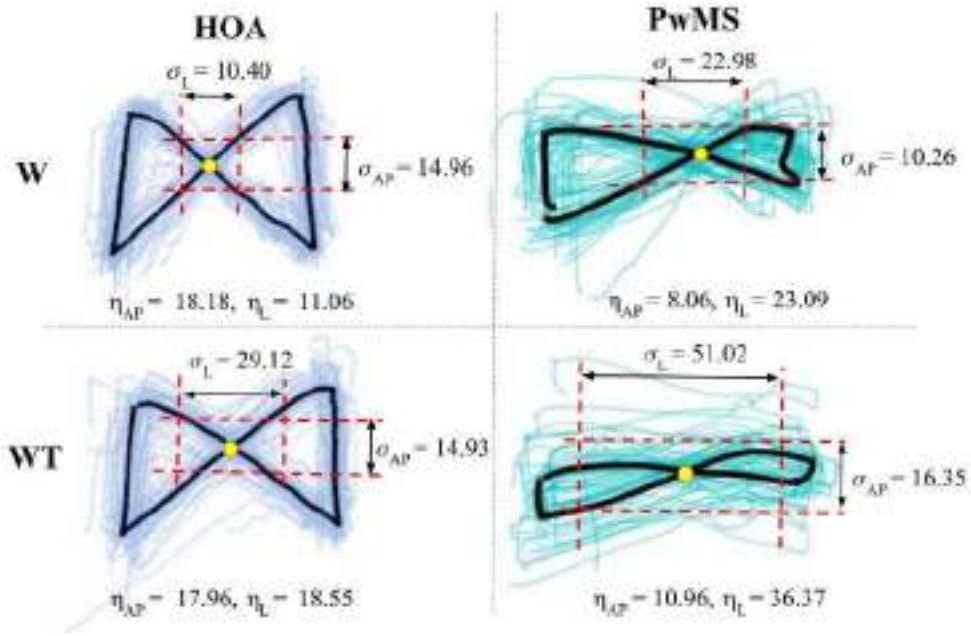

Figure 2.2: Butterfly diagram. Left: HOA, Right: PwMS with EDSS = 5.5, Top: Trial W, Bottom: Trial WT. The curves illustrate the butterfly diagram during the entire 75 seconds walk where the thicker black line and yellow circle depicts the mean trajectory and mean intersection point, respectively. Here, x and y axis represent the center of pressure position coordinates. The variability (red dashed lines) and asymmetry in the anterior-posterior (σAP /ηAP) and lateral (σL/ηL) directions are reported in mm.

2.2.2.3 Datanormalizationtechniques The demographic differences between subjects may intrinsically influence the dynamics of gait
variability and hence bias the MS gait differentiation efficiency. Thus, prior to classification, we normalized the subject's derived gait characteristics using the following two approaches:

• Body size-based dimensionless normalization (Size-N): The extracted gait variables were normalized to non-dimensional forms by dividing via their corresponding dimension-matched body size-based scaling factors (proposed in [110]) in order to adjust for the inherent inter-subject physical differences. For instance, the acquired lengths, namely stride length and width were scaled by the subject's respective height. FPAs are dimensionless and thus require no scaling. Let w, h, Ssize and g denote the body weight (in kg), height (m), shoe size (m) and acceleration of gravity (9.81 m/s2), respectively, then Table 2.1 summarizes scaled dimensionless quantities with regards to features obtained for both cohorts and trials where L ∈ {SL, SW}, T ∈ {SSR, DSR, DSL, ST, StT, SwT}, Fez ∀ e ∈ E, θ ∈ {θL, θR} and P ∈ {βL, αL}.

Table 2.1: Size-N normalization for the extracted gait features

Raw feature	L	T	Fez	C	SS	θ	W	P
Scaled feature	$\frac{L}{h}$	$\frac{\sqrt{T}}{\sqrt{\frac{h}{g}}}$	$\frac{Fe}{wg}$	$\frac{C}{60\sqrt{\frac{g}{h}}}$	$\frac{SS}{\sqrt{Sg/h}}$	θ	$\frac{W}{60\sqrt{hg}h}$	$\frac{P}{S_{size}}$

• Multiple regression-based normalization (Regress-N): Gait variables from both walking trials of the 35 subjects (in 2.2.1.2) were normalized by regressing the baseline gait features of normative walking data from 30 additional healthy older adults on multiple demographic characteristics. These additional healthy older adults (age: 67.6 ±10.34 years [50–87 years], weight: 71.61 ±14.52 kg [52.97–103 kg], height: 1.68±0.17 m [1.01–1.96 m], male/female: 9/21) were recruited from the local community. All controls walked for 200 s on the treadmill and yielded 21 gait features from a total of 3923 valid strides. A regression model was fitted to each gait feature with subject-wise averaged gait parameter values as a dependent variable and their corresponding demographics (weight, height, gender and age) as independent variables. In other words, Featureregression normalized = Featureraw/L(demographics). All independent variables were used while fitting the regression since the variance inflation factor for each was lower than 5 [111], hence ignoring the concern of multicollinearity. Further, the Spearman's rank correlation coefficients among

the independent variables presented no strong associations. For each gait feature, backward elimination was used to determine M statistically significant predictors ($p < 0.1$) and an optimal combination of predictors with the minimum corrected Akaike informa- tion criterion was selected out of 2M possibilities. Subsequently, robust regression models minimizing the Tukey's biweight loss of the standard Gaussian residual errors were fit (see Table A.2 in the for the regression coefficients and the corresponding root mean squared errors). Gait features from both trials of the 35 study subjects (in 2.2.1.2) were then normalized to dimensionless quantities with their predicted values obtained via their corresponding fit and subject demographics. Scaling relative to the regression predictors and coefficients computed from normative walking data of other healthy older adults aids in minimizing data spread among the gait features for the controls and association with individual demographic characteristics, and thus improve detection of MS vs. subject-related changes in gait.

2.2.2.4 Statisticalanalysis

To examine cohort-related differences and the corresponding effect of normalization strategies on gait feature characteristics (i.e. mean, standard deviation and range), we used a two tailed t-test and F-test to identify significant MS-related differences at $\alpha = 0.05$. The statistical assumptions of independence (since all subject observations were independent), normality (via the Shapiro Wilk test) and homoscedasticity (via Levene's test) were verified for the t-test. Mann-Whitney U-test and Welch's t-test were used, respectively, if normality or homoscedasticity, respectively failed. Similarly for the F-test, independence and normality were examined, and Levene's test was implemented if normality failed. Spearman's correlation (r) between the mean gait parameters and physical characteristics (weight, age, height and gender) of subjects in both trials were compared for raw (rraw), body size (rs), and regression (rreg) normalized data to study the dependence of gait features (and thus the performance of machine learning models) on subject demographics. Further, among PwMS, we explored the association and directionality of raw and normalized gait variables with disease severity using Spearman's correlations (redss) in order to motivate the applications of gait in learning MS progression with time.

2.2.2.5 Classification models and evaluation

MS prediction was studied across two classification designs, namely task and subject generalization (Figure 2.1). In both task and subject generalization, binary supervised learning classifiers were trained to differentiate strides corresponding to HOA and PwMS. Machine learning mod-

els were trained on 1654 strides across all 35 subjects in W trials and tested to categorize 1576 strides of the same subjects in WT trials for task generalization. Since our data set was limited to 35 subjects, we used a 7-fold cross-validation for subject generalization. In each scenario, all models were examined with both size-N and regress-N normalized features. Z-score normalization was applied to all features to eliminate the influence of variable feature ranges. For both classification architectures, the performance of nine notable supervised classifiers, i.e. decision tree (DT), random forest (RF), support vector machine with linear (LSVM) and radial basis func- tion (RBF SVM) kernels, gradient boosting machine (GBM), adaptive boosting (AdaBoost), eX- treme gradient boosting (XGBoost), multilayer perceptron (MLP) and logistic regression (LR) were compared (see Appendix Section A.1.2.1 for details on these algorithms). Exploratory hyperparameter optimization was performed to determine optimal framework for each learning classifier. Prediction efficiency for the task and subject generalization classifiers were weighed via the test set and mean cross-validation precision, recall, accuracy, F1 score and area under receiver operating characteristic (ROC) curve (AUC) metrics, respectively. Both setups were evaluated at stride and subject level, where majority voting was used to classify subjects into HOA vs. PwMS. Thus, a correctly classified subject's walk had more than 50% of strides accurately detected as of the appropriate cohort. Precisely, we annotate the stride and subject-level classification metrics with str (i.e. Pstr, Rstr, Astr, F1str, AUCstr) and sub (i.e. Psub, Rsub, Asub, F1sub, AUCsub) subscripts, respectively.

2.2.2.6 MSprogressionspace We attempt to describe the progression stage in PwMS by clustering their strides in distinct
and multifaceted progression subgroups. Dimensionality reduction via rank-2 non-negative matrix factorization (NMF) was implemented on 21 regress-N features with 749 and 698 available strides of PwMS in trials W and WT, respectively to define a progression space for MS summarizing the influence of gait features in 2 dimensions (2D) across multiple stages. To impose non-negativity, all regress-N features were normalized between 0 and 1. Across both trials, NMF decomposed the data into two matrices, which we interpret as progression vectors and the progression indicators. Progression vectors were used to construct the 2D MS progression space (2D-MSPS). The 21 gait features were correlated with the two axes of the progression space using the magnitude of coefficients observed in the progression indicator vectors. Next, by applying unsupervised Gaussian mixture model (GMM) on the 2D-MSPS, we algorithmically parsed the progression space into three hidden subtypes within PwMS, representing the disease rate progressors. GMM is an expectation-maximization algorithm that maximizes the likelihood of observing the data, given the underlying parameters of the distribution. For each identified

cluster, we study the number of strides and their share percentage in three severity subgroups (defined in 2.2.1.2) based on the EDSS of MS subjects. Further, we look at the weights of the features to define a projection mapping for gait variables to the new 2D MSPS axes and thus find latent features describing the reduced progression space.

2.3 Results

Overall, PwMS reported longer and more dispersed stride time, stance time and double support times but a shortened single support time on average in both the trials. Further, PwMS walked with a reduced stride length, cadence, self-controlled speed and a wider lateral distance between the two feet. PwMS reported higher median and spread in the butterfly diagram extracted lateral shift (βL) and squared deviation (αL). In general, no individual or combination of features exhibit clear non-overlapping patterns characterizing MS. Any statistical model for MS prediction would thus be very high dimensional and prone to substantial scale and validation challenges. Therefore, machine learning is an appropriate approach for the MS identification task.

2.3.1 Statistical Analysis

2.3.1.1 Statisticalsignificance

Subject-wise averaged raw and normalized features were compared between HOA and PwMS for significance of difference in means and variances. Considering trial W, statistically significant differences between means were observed in raw left FPA (6.4 times higher (6.4 × in PwMS) and squared deviation (1.9× in PwMS). After the body size-normalization, terminal double support (1.6 × in PwMS), force on TOL (1.5× in PwMS), lateral shift (1.7 × in HOA), lateral shift (1.6× in PwMS) and squared deviation (1.9× in PwMS) based on size-N, terminal double support (1.4 × in PwMS), lateral shift (1.6× in PwMS) and squared deviation (1.8 × in PwMS) were significant and using the size-N data, terminal double support (1.5 × in PwMS), lateral shift (1.6× in PwMS) and squared deviation (2.7× in PwMS) exhibited statistical significance. Similar to size-N, regress-N terminal double support (1.5 × in PwMS), lateral shift (1.6× in PwMS) and squared deviation (2.6× in PwMS) showed significance in trial WT. 8 raw, 10 size-N and 12 regress-N features in trial W and 11 raw, 14 size-N and 14 regress-N features in trial WT indicated significant differences between variances.

In essence, both the body size- and regression-based normalization increased the number of parameters that exhibit significant difference between means and variances of the two cohorts.

2.3.1.2 Correlationwithphysicalfeatures

To explore the dependency of gait features on demographics, correlation (r) of physical properties with raw (rraw), size-N (rs) and regress-N (rreg) parameters was compared. Across both trials, the range of correlations with raw data (W: $-0.41 \le r_{raw}^{W} \le 0.91$, WT: $-0.46 \le r_{raw}^{WT} \le 0.89$) lowered with size-N ($-0.46 \le r_{s}^{W} \le 0.56$, $-0.49 \le r_{s}^{WT} \le 0.53$) and further declined with regress-N features ($-0.41 \le r_{reg}^{W} \le 0.41$, $-0.44 \le r_{reg}^{WT} \le 0.51$). Figure 2.3 plots some of these absolute correlations for trial W. For instance, size-N toe-off forces demonstrated significantly weaker correlations ($0.13 \le |rs| \le 0.22$) with subject's height than their raw counterparts ($0.4 \le |rraw| \le 0.43$). A similar trend was observed for the heel strike forces as well along with a further decrease for regress-N forces. High correlations between raw forces and subject's weight ($0.81 \le |rraw| \le 0.91$) and gender ($0.42 \le |rraw| \le 0.62$) weakened considerably with size-N to $0.03 \le |rs| \le 0.46$ and $0.12 \le |rs| \le 0.19$, respectively and with regress-N forces to $0.01 \le |rreg| \le 0.51$ and $0.01 \le |rreg| \le 0.41$, respectively. Interestingly, interaction between single support and gender heightened from 0.1 $\le |rraw| \le 0.3$ to $0.24 \le |rs| \le 0.41$ with size-based normalization. Regress-N weakened ($|rreg| \le 0.25$) most associations with very infrequently realizing moderate values ($0.25 < |rreg| \le 0.51$) over both trials. Specifically, prominent correlations between weight and stride width ($|r_{raw}^{W}| = 0.61$, $|r_{raw}^{WT}| = 0.62$), left FPA ($|r_{raw}^{W}| = 0.46$) and right FPA ($|r_{raw}^{W}| = 0.24$, $|r_{raw}^{WT}| = 0.52$) distinctly lowered to $0.02 \le |rreg| \le 0.32$. Size-N could not assist in diminishing these high associations between stride width, left/right FPA and weight. All high correlations ($|r| \ge 0.7$) reduced to moderate ($0.5 \le |r| \le 0.7$) or low ($|r| \le 0.5$) values with normalization. Thus, normalization reduced the inherent subject specific differences associated with physical characteristics in the gait features, potentially enabling the machine learning models to focus on learning to differentiate only disease-specific characteristics present in the gait parameters and consecutively increase their test set generalizability.

2.3.1.3 Correlationwithdiseaseseverity

To explore the association of gait parameters with severity among PwMS, correlation (redss) of EDSS with raw and normalized features was studied. Figure 2.4 plots the correlations for trial W. The directionality of redss matched our instinct with speed, length and cadence inversely correlating; and stride, stance, double support times and lateral shift positively interacting with disability. With respect to all three data streams, EDSS showed the strongest negative corre-

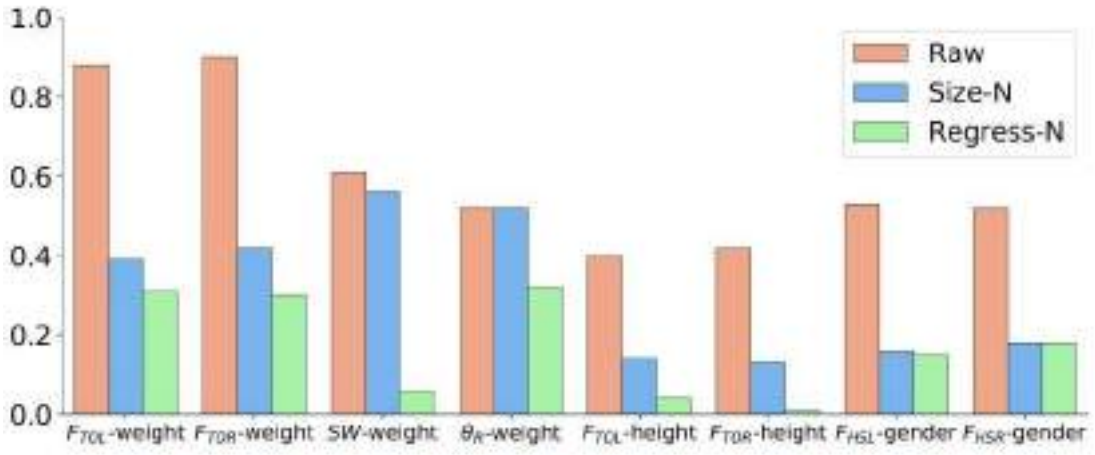

Figure 2.3: Correlation with demographics.Absolute correlation of raw (red), size-N (blue) and regress-N (green) features with physical characteristics in trial W.

lations (redss ≥ 0.7) were illustrated with double support and stance times in both trials and also with stride time in trial WT. Cadence in trial W and walk ratio in trial WT exhibited moderate negative associations (−0.7 < redss ≤ −0.5). Moderate positive interactions (0.5 ≤ redss < 0.7) were shown by lateral shift and only normalized forces at MidSSR in both trials as well as stride time in trial W and lateral deviation, force at MidSSL in trial WT. The computed correlations were statistically significant for nine raw and normalized parameters (SL, SS, C, W , ST , StT , DSI, DST and αL) in both trials and two additional variables (FMidSSL and βL) in trial WT. The correlation of forces at MidSSR demonstrated significance only after normalization. Significant correlations between gait characteristics and EDSS motivate the applications of gait in learning the progression space and clinical stages of MS.

2.3.2 Prediction Models

Nine classifiers were compared with size-N and regress-N data to categorize strides and subjects between HOA and MS cohorts for task (2.3.2.1) and subject (2.3.2.2) generalization.

2.3.2.1 Taskgeneralization

To examine the differences of single and dual-task walking on individual gait characteristics in older adults with and without MS, we used a linear mixed effects model. Overall, all individuals

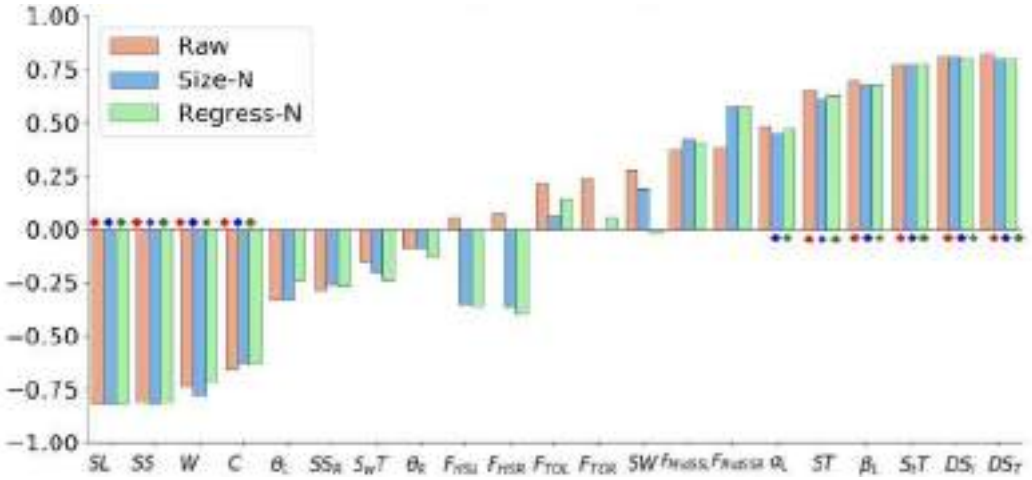

Figure 2.4: EDSS Correlation. Bar plot illustrating the correlation of raw (red), size-N (blue) and regress-N (green) features with EDSS in trial W. Statistically significant correlations are marked with diamonds of respective colors.

demonstrated a significant increase in stance time, initial and terminal double supports and forces at MidSSR and TOR, and a significant decrease in stride length and speed when going from W to WT trials. A significant two-way interaction between cohort and task indicates greater increases in stride, stance, swing and right single support times, stride length, speed and walk ratio for PwMS during WT trials compared to HOA during W trials. A significant decrease in stride width, cadence and forces at HSR, TOL, MidSSR and HSL was observed for PwMS in WT compared to HOA under W trials.

Table 2.2 summarizes the stride- and subject-wise evaluation metrics for top-5 task generalization classifiers on categorizing the test set strides of trial WT (see Table A.3 in the Appendix for hyperparameter exploration). Clearly, aggregated performance of all the subject's strides via majority voting improved upon the accuracy of individual stride-wise predictions, for instance from 74.3% to 82.9% and 79.2% to 94.3% on RF with size-N and regress-N data, respectively. The classification performances of all algorithms were higher across all metrics with the regress-N data except only for GBM with equal subject-wise metrics when using the size-N and regress-N data. LR, DT, linear SVM and AdaBoost are absent from table 4.2 of top-5 classifiers. RF, RBF SVM and GBM achieved a subject classification accuracy (Asub) of 94.3% with the regress-N data while Asub for XGBoost and MLP were 91.4% and 88.6%, respectively with the regression normalized data. RF, RBF SVM, XGBoost and MLP resulted in an Asub of less than 90% with the size-N data except GBM that matched the 94.3% accuracy of regress-N. The maximum stride classification AUC (AUCstr) was 0.91 followed by 0.90 using the regress-N data on GBM and XGBoost, respectively whereas the optimal AUCstr with the size-N data was 0.87 on GBM and XGBoost. RF, RBF SVM and MLP had an AUCstr of less that 0.85 when using the size-N data.

Table 2.2: Task generalization: stride- and subject-wise test set performance for top-5 algorithms

		Stride-based					Subject-based				
Algorithm	Data	Accuracy	Precision	Recall	F_1	AUC	Accuracy	Precision	Recall	F_1	AUC
RF	Size-N	0.743	0.73	0.666	0.69	0.841	0.829	0.923	0.706	0.80	0.935
	Regress-N	0.792	0	0.713	7	0.886	0.94	1.0	0.882	0.938	0.987
RBF SVM	Size-N	0.744	0.79	0.779	0.75	0.819	3	0.833	0.882	0.857	0.980
	Regress-N	0.785	6	0.841	2	0.868	0.941	0.941	0.941	0.941	0.997
GBM	Size-N	0.787	0.68	0.716	0.73	0.867	0.85	1.0	0.882	0.938	1.0
	Regress-N	0.824	6	0.729	0	0.91		1.0	0.882	0.938	1.0
XGBoost	Size-N	0.784	0.72	0.732	0.77	0	7	0.933	0.824	0.875	0.980
	Regress-N	0.815	0	0.735	6	0.867			0.824	0.903	
MLP	Size-N	0.746	0.78	0.652	0.74	0.901	0.94	1.0	0.706	0.80	1.0
	Regress-N	0.795	4	0.648	9	0.820	3	0.765	1.0	0.867	
			0.85		0.78	0.878		0.923			0.951
			3		6		0.94	1.0			0.974
			0.77		0.75		0.94				
			0		0						
			0.82		0.77		3				
			7		8		0.88				
			0.74		0.69		6				
			2		4		0.91				
			0.85		0.73		4				
			4		7		0.82				
							9				
							0.88				
							6				

Note: AUC is area under the receiver operating curve, Size-N is body size-based normalization, and Regress-N is multiple regression-based normalization. The numbers in bold represent the highest model performance.

Considering all evaluation metrics in table 4.2, GBM with regress-N data performed the best with an accuracy, F1 and AUC of 82.4%, 0.79 and 0.91, respectively at stride-level and 94.3%, 0.94 and 1.0, respectively at subject-level classification, followed by RF and RBF SVM on regress-N with a matching subject-level accuracy. Boosting algorithms sequentially optimized the current DT by adapting to the errors on the data of prior weak learners as compared to RF training DTs in parallel on bootstrap samples, thus GBM significantly improved the performance of learners with low variance but high bias. Gradient boosters iteratively regress over negative gradients of any generic differentiable loss function to boost the weak learning DTs whereas AdaBoost reweighing the previously mistaken data points higher specifically optimizes an exponential loss. MLPs are efficient to form disconnected decision regions and learn any arbitrary complicated boundary, as suggested by the universal approximation theorem. The optimal task generalization algorithm was GBM trained on regress-N data with 150 boosting stages, depth of 7, learning rate of 0.15 and considered 5 features for checking the best split (see Figure A.7 for its confusion matrix). Only two PwMS were miss-classified as HOA.

2.3.2.2 Subjectgeneralization

Table 2.3 summarizes the mean and standard deviation of 7-fold cross-validation performance metrics for the top-5 subject generalization classifiers (see A.4 for optimal hyperparameters).

All algorithms except AdaBoost with regression normalization surpassed the diagnostic performance when using the standard size-based normalization. LR, linear/RBF SVM and XG-

Table 2.3: Subject generalization: stride- and subject-wise mean cross-validation performance for top-5 algorithms

		Stride-based					Subject-based				
Algorithm	Data	Accuracy	Precision	Recall	F_1	AUC	Accuracy	Precision	Recall	F_1	AUC
DT	Size-N	0504 012.	050025.	0459 020.	0427 015.	0538 012.	0514 034.	0429 043.	0429 042.	0410 039.	0690 038.
	Regress-N	0541 008.	0526 022.	0512 019.	0467 011.	0597 011.	060024.	0476 038.	050038.	0462 034.	0679 027.
RF	Size-N	0533 016.	0547 028.	0418 024.	0408 019.	0635 023.	0571 025.	0548 034.	0548 033.	0514 029.	069027.
	Regress-N	0563 011.	0557 025.	0463 023.	0449 016.	0643 016.	060019.	0571 032.	0524 027.	0519 024.	0643 019.
GBM	Size-N	0538 018.	0557 029.	0453 025.	0434 020.	0617 022.	0486 028.	0333 036.	0429 042.	0371 038.	0726 029.
	Regress-N	0584 012.	0580 024.	0518 023.	0486 017.	0654 014.	060024.	0452 033.	050038.	0471 035.	0798 020.
AdaBoost	Size-N	0592 015.	0595 ± 023	0440 024.	0451 019.	0644 018.	0543 018.	0429 032.	0357 023.	0381 025.	0774 015.
	Regress-N	0586 010.	0562 028.	0432 019.	0459 017.	0598 019.	060021.	0524 038.	0452 033.	0467 032.	0631 030.
MLP	Size-N	0524 014.	0534 028.	0362 024.	0366 019.	0598 020.	0571 020.	0524 038.	0405 032.	0424 029.	0762 028.
	Regress-N	0621 ± 010	0579 022.	0619 ± 020	0565 ± 014	0682 ± 015	080 ± 015	0833 ± 020	0786 ± 025	0776 ± 017	0857 ± 023

Note: The numbers in bold represent the highest model performance.

Boost did not make it to top-5. The best mean Asub was 80% (95% confidence interval (CI): [75, 85]) using the regress-N data with MLP while RF and MLP had the maximum Asub of 57.1% with the size-N data. Overall in table 4.3, MLP with regress-N data performed the best with a mean accuracy, F1 and AUC of 62.1%, 0.57 and 0.68, respectively at stride-level and 80%, 0.78 (95% CI: [0.72, 0.83]) and 0.86 (95% CI: [0.78, 0.93]), respectively at subject-level classification. Tree-based models handle highly correlated variables to avoid overfitting better than kernel SVM. Unlike traditional machine learning algorithms relying wholly on hand-crafted features, MLPs are capable of incrementally learning latent characteristics of the data and discover novel inherent feature hierarchies with increasing complexity of the design. Our optimal MLP architecture with 7 fully connected layers and ReLU non-linearity was trained for 200 epochs using the adaptive moment estimation (Adam) optimizer with an adaptive learning rate initially set to 0.001 and the cross entropy loss (see Figure A.7 for its confusion matrix). Four PwMS and three HOA got incorrectly classified. Thus, GBM achieved the best Asub (94.3%) for task generalization, whereas MLP performed the best (80%) for subject generalization.

2.3.3 Post hoc Analysis

Note that for further analysis, we adhered to only using regress-N data for it demonstrated superior performance across both task and subject generalization model designs.

2.3.3.1 Ablationstudy

We compared the task and subject generalization performance on several subsets of regress-N features, namely 4 spatial (S), 7 temporal (T), 8 kinetic (K), 13 spatiotemporal (ST), 12 spatial-kinetic (S+K) and 15 temporal-kinetic (T+K) parameters, to that of using all 21 variables for MS prediction. All machine learning models were tuned from scratch on these data streams for comparison. Table 2.4 illustrates the subject-wise metrics for the best performing algorithm on each subset across both the task and subject generalization schemes. Across both model de-

Table 2.4: Ablation study: Task and subject generalization models

	Taskgeneralization					Subject generalization				
Data	Topalgorithm	A_{sub}	P_{sub}	R_{sub}	F_{1sub}	AUC $_{sub}$	Topalgorithm	A_{sub}	P_{sub}	R_{sub}	F_{1sub}	AUC $_{sub}$
Spatial	RF XGBoost	0.74	0.86	1	0.93	0.91	GBM MLP MLP RBF SVM AdaBoost AdaBoost MLP	0.66 ± 0.21	0.71 ± 0.22	0.33 ± 0.45	0.45 ± 0.29	0.61 ± 0.26
Temporal	GBM GBM	0.80	0.92	0.91	0.91	0.92		0.69 ± 0.21	0.69 ± 0.33	0.63 ± 0.37	0.61 ± 0.31	0.77 ± 0.25
Kinetic	RBFSVM	0.80	0.94	1	0.97	0.98		0.63 ± 0.20	0.57 ± 0.49	0.26 ± 0.23	0.36 ± 0.31	0.56 ± 0.16
Spatiotemporal		0.83	1.0	0.91	0.95	0.98		0.71 ± 0.18	0.81 ± 0.25	0.51 ± 0.35	0.58 ± 0.30	0.71 ± 0.25
Spatial-kinetic		0.80	1.0	0.88	0.93	0.98		0.71 ± 0.21	0.76 ± 0.24	0.63 ± 0.37	0.64 ± 0.31	0.80 ± 0.23
Temporal-kinetic	MLP	0.91	1.0	0.94	0.97	0.98						
All	GBM	0.94	1.0	0.94	1.0	1.0		0.80 ± 0.15	0.83 ± 0.20	0.78 ± 0.25	0.77 ± 0.17	0.85 ± 0.23

Note: The numbers in bold represent the highest model performance.

signs, LR, DT and linear SVM were never the top performers. Overall, GBM and MLP followed by AdaBoost are the most prominent algorithms in Table 2.4 for task and subject generaliza- tion, respectively. Task generalization revealed the best performance when using all 21 features with GBM (Asub: 0.94, AUCsub: 1.0) followed by spatiotemporal also with GBM (Asub: 0.94, AUCsub: 0.98) and temporal-kinetic parameters with MLP (Asub: 0.91, AUCsub: 0.98). For subject generalization, MLP with all features had the best mean results (Asub: 0.80, AUCsub: 0.86) followed by temporal-kinetic with AdaBoost (Asub: 0.71, AUCsub: 0.80) and spatial-kinetic also with AdaBoost (Asub: 0.71, AUCsub: 0.71). In both model designs, machine learning algorithms had a better performance using all features, thus these ablation results indeed support our decision to use all the extracted gait features for prediction.

2.3.3.2 Analysisoffeatureimportance

We first investigated the importance of features via conditional entropy (CE). The CE of labels Y, taking binary values, with respect to the discretized feature X, taking values in a finite set X, was defined as: $\sum_{(x,y)\in X\times\{0,1\}} p_{X,Y}(x,y)\ln\frac{1}{p_{X,Y}(x,y)} - \sum_{x\in X} p_X(x)\ln\frac{1}{p_X(x)}$, where $p_{X,Y}$ joint probability mass function of (X,Y) and p_X is the probability mass function of X. Features with a low entropy reflect less randomness and hence are more predictive of labels. Figure 2.5 depicts the CE of all features in trials W and WT. The most informative features with the least CE

were (in order) SL > SS > C > FTOL > SwT in trial W and SS > FTOR > FTOL > SL > W in trial WT. Cadence followed by swing time in trial W and terminal double support followed by stance time in WT showed the most reduction in entropy among temporal features. Stride length followed by width from spatial, stride speed out of spatiotemporal and toe-off forces from kinetic features delivered the most predictive power in both trials. Overall, stride speed, length and forces at the toe-off were found to be the most valuable features across both trials. FPAs and lateral deviation with a high CE in both trials were least predictive of the labels. Given that

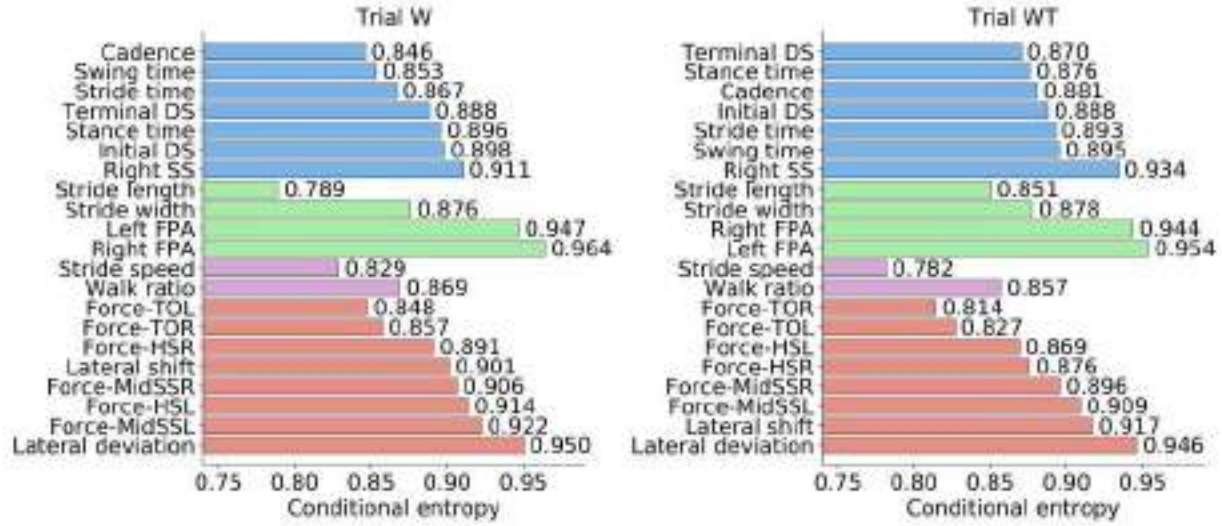

Figure 2.5: The entropy present in the labels given regress-N gait features in trials W (left) and WT (right). Temporal, spatial, spatiotemporal and kinetic features are grouped in blue, green, plum and red colors, respectively.

our best machine learning algorithms, GBM and MLP for task and subject generalization, respectively, used all 21 features, we also investigated feature importance by studying the decrease in performance of optimally tuned GBM and MLP models when only including features from specific subsets. Apart from subsets S, T, K, ST, S+K and T+K considered in Section 2.3.3.1, we defined another group as features obtainable from wearable sensors for this analysis. All defined gait features except the butterfly diagram-based parameters could be derived from wearable foot switches or inertial sensors [112]. Figure 2.6 depicts the AUCsub for optimal task (GBM) and subject (MLP) generalization models with features from several data domains. For both models, using all features yielded the best AUCsub, followed by wearable-derivable measures (0.998) and spatiotemporal (0.977) features for task generalization and by spatiotemporal (0.738) and wearable-derivable/kinetic (0.726) parameters for subject generalization. In both frameworks, no one set of features outperformed or matched the performance of using all features collectively. Especially for subject generalization, all features together are essential to diagnose the

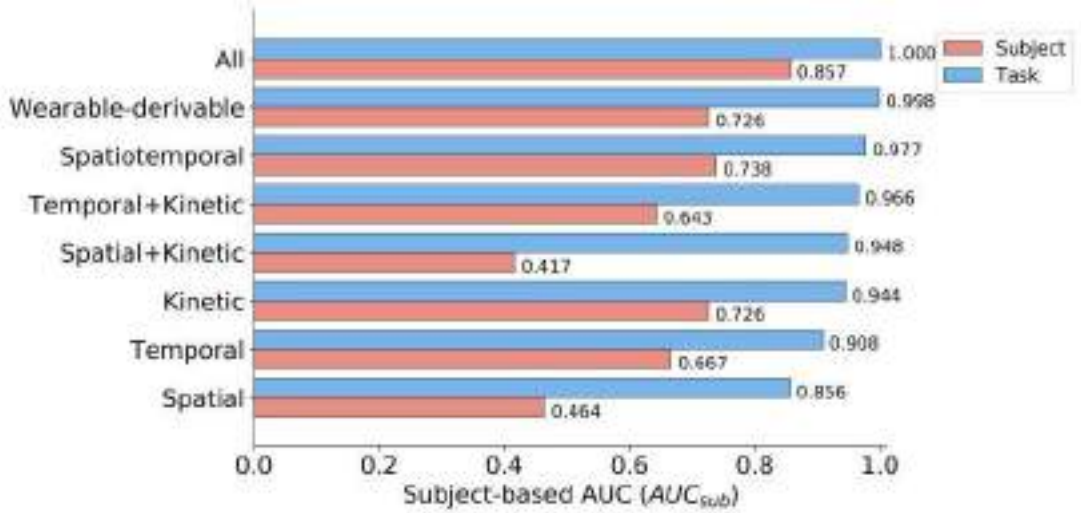

Figure 2.6: Feature importance. AUCsub for task and subject generalization models with different data domains are represented in blue and red, respectively.

heterogeneity present in new subjects.

2.3.4 MSProgressionSpace

Promising correlations between gait features and EDSS (Section 2.3.1.3) motivated exploring gait-based characteristics to describe the MS progression space. To define hidden clinical sub- types within PwMS, unsupervised GMM was used to partition the NMF reduced 2D-MSPS. In both trials, three optimum number of underlying clusters for GMM were attained using the Bayesian information criterion (BIC). Figure 2.7 depicts the three identified clusters in strides of PwMS with distribution in strides of controls superimposed for visualization in both trials. For each identified cluster, Table 2.5 summarizes the number of strides and their share percentage in three severity subgroups based on the EDSS of MS subjects. Cluster 1 (green) is dominated by

Table 2.5: Count and ratio of strides relative to EDSS in each cluster. Clusters 1, 2 and 3 are abbreviated as C1, C2 and C3, resp.

	TrialW			Trial WT		
EDSS	C1	C2	C3	C1	C2	C3
1.0-2.5 (mild)	131 (0.29)	17 (0.07)	0 (0.0)	149 (0.31)	12 (0.07)	0 (0.0)
3.0-4.5 (mild-to-moderate)	317 (0.69)	23 (0.09)	0 (0.0)	308 (0.65)	11 (0.06)	0 (0.0)
5.0-6.0 (moderate)	11 (0.02)	204 (0.84)	46 (1.0)	18 (0.04)	148 (0.87)	52 (1.0)

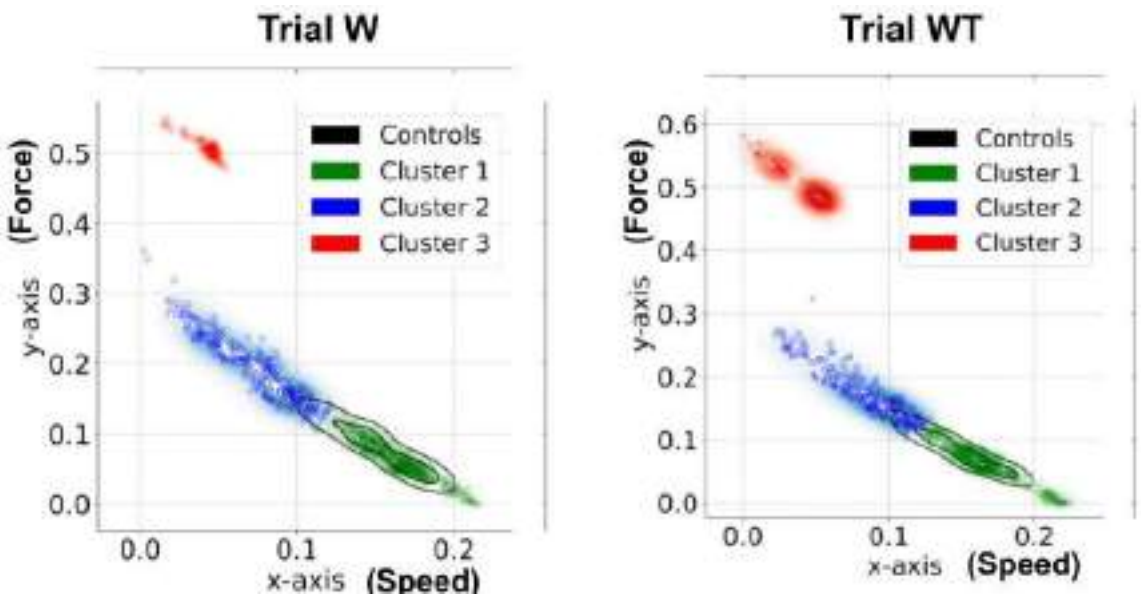

Figure 2.7: Constructed 2D-MSPS. Left: Trial W, Right: Trial WT. Three clusters (shown in green, blue and red) are identified in strides of PwMS and distribution of HOA strides is depicted in black outlines.

strides of mild and mild-to-moderate severity patients. Cluster 2 (blue) is majority of moderate PwMS strides covering around 84% in trial W and 87% in WT of cluster observations and clus- ter 3 (red) has no mild or mild-to-moderate strides and contains only strides of moderate PwMS. The share of mild and mild-to-moderate strides is decreasing with an increase in the progression rate. Visually, distribution of control strides most overlaps with cluster 1 dominated by strides from mild and mild-to-moderate subgroups. Further, we looked at the weights of the 21 features to define a projection mapping for gait variables to the new 2D MSPS axes (see Figure A.8 in the Appendix). For both trials, the horizontal axis was dominated by stride speed and its related components and vertical axis corresponded to force related features. Interestingly, gait speed and force measures were the top predictive power features too (as found in Section 2.3.3.2).

2.4 Discussion

This study examined MS and disability related changes in spatiotemporal and kinetic gait features after normalization; and evaluated the effectiveness of GML4MS to classify strides of PwMS from healthy controls, and generalize across different walking tasks and subjects after gait normalization. A few other works have explored machine learning to classify MS based on gait data. Gait features extracted from 3D ground reaction force data were adopted to dis-

criminate healthy, cerebral palsy and MS subjects using two machine learning methods, namely nearest neighbours and MLP [96]. However, a very modest dataset with only four PwMS was employed for this study and thus limits the generalization of the classification results. Further, the study is limited in examining only force data and not exploring any tree-based machine learning algorithms. A recent study used smartphone and smartwatch sensors data and machine learning to distinguish among healthy controls, mildly (PwMSmild) and moderately (PwMSmod) disabled PwMS during a two-minute walk test [97]. Although this work investigates three well-known algorithms, namely, LR, SVM and RF to achieve the best accuracy of 82% differentiating PwMSmod from HOA and from PwMSmild and 66% identifying PwMSmild from HOA; the analysis on boosting algorithms, which have known to outperform RF in most applications, is missing. Moreover, our study utilizes up to 75 seconds of data for analysis, as compared to the longer data sample of two-minute walk in [97]. Another recent work analyzed a long short-term memory approach to classify fall risk in PwMS using accelerometers [113]. To the best of our knowledge, this is the first study utilizing data driven machine learning for classification of individual strides of older PwMS using both spatiotemporal and kinetic features while walking. Our stride-based feature extraction approach derived multiple samples from a single subject, thus augmenting and introducing significant variations to our dataset to improve the generality of machine learning classifiers, which may allow for frequent and even real-time inferences.

The instrumented treadmill adopted for this study allowed for continuous gait monitoring of longer durations and distances within a compact footprint, relative to overground walking, and the capture of deviations from several successive strides [114]. While PwMS in this study were able to walk independently, the ceiling mounted harness, rails, and emergency stop provide essential tools for safety in PwMS with balance and fatigue concerns. Further, the integration of a built-in force plate supported kinetic data acquisition and allowed for online detection of gait events [105]. While walking on a treadmill can affect gait performance [115], these differences are generally within the normal variability of gait parameters and may be further diminished after an appropriate accommodation period to treadmill walking [116]. Our treadmill training before actual data collection and adaptive speed control helped subjects to more closely resemble natural walking.

Our work examined the benefits of regression normalized gait features on the accuracy of MS prediction using stride-based machine learning classification algorithms. Both the size- and regression-based normalization schemes increased the number of parameters demonstrating statistical significance between HOA and PwMS. The ability of regress-N normalization to reduce the association between gait features and personal demographics is crucial towards boosting the performance and generalizability of machine learning classifiers aimed at MS prediction. We have used statistical insights from admittedly a small number of test subjects. However, through

the extraction of regress-N gait features, our approach mitigates some of the concerns related to small sample sizes since we are reducing the bias in the data by increasing independence (see Section 2.3.1.2). Compared to past studies on regression normalization in machine learning for other neurological disorders [95, 99] using the same controls in their classification set to extract regression coefficients, we used a normative dataset separate from our 35 study subjects to derive regression models for the gait features, hence prohibiting any divulgence of information from validation to training set.

Our proposed task generality framework demonstrates the feasibility of training on data col- lected in a lab-based walking task, and prediction on a walking-while-talking task, which paves the way for further inquiry into prediction using data collected in naturalistic and ecologically valid scenarios. We conclude that regress-N data with GBM and MLP were the optimal ma- chine learning frameworks for task and subject generalization, respectively. An ablation study on the set of features supported using all the extracted gait features for better predictability in both model designs. From a clinical perspective, stride level classification allows for the use of a single stride, or brief duration walking trial, to serve as the basis for disease progression monitoring, which may be well suited for clinical settings with limited space and time. Further, as an effort towards the explainability of our ML-based study, we explored conditional entropy and decreases in performance of optimal GBM and MLP models. When only including a subset of features to examine the most relevant features driving the machine learning performance, we found that stride speed, length and forces at the toe-off were the most valuable features across both trials. Furthermore, we find that the use of wearable-derivable features is closely behind all features in terms of classification performance, which provides preliminary evidence of the feasi- bility of using wearable sensor data collected at home or local community in future telemedicine or rural health applications. Our study also examined how well normalized gait features could predict disability in PwMS. Significant correlations between gait characteristics and disability in PwMS (see Section 2.3.1.3) motivated the application of regress-N gait features in learning the progression space of MS (see Section 2.3.4). Of particular significance, the two reduced dimen- sions arising after NMF were dominated by stride speed and force, which were also the most predictive features of MS-related changes.

The current work designs a domain knowledge-based MS screening model but the small cohort size recruited for this study limits making generalized interpretations for the heterogeneous MS community. Although, the features selected for predictive models in this study, namely, spatiotemporal characteristics (see [86–90]), FPAs [108], butterfly diagram-based variables [86,109] and forces [91], have been clinically shown and commonly adopted in the past to quantify gait impairments in PwMS, yet, by pre-selecting a specific set of domain knowledge-based features, we might be at a risk of introducing certain investigator bias in our machine learning models.

Future work should focus on carefully characterizing the potentially missed information represented by the non-selected variables. Machine learning explainability analysis in Section 2.3.3 serves as an initial estimate to demonstrate the influence of our feature selection on the model prediction performance. For an ideal understanding of dynamics from the inherently continuous gait data stream [117], we would need further exploration on non-linear dynamical features characterizing the human movement. Future research should examine associations of gait parameters with additional demographic and clinical factors to design improved normalization techniques. Further evaluation of GML4MS on a separate MS dataset with additional concurrent tasks, or while walking at home or in the community would be essential to establish robustness and im- prove sensitivity. Exploring hidden Markov and recurrent neural network predictive models by using tensors of independent strides will be vital to gauge the temporal component present in the continuous gait data. Future work is needed to identify prospective fall risk in MS subjects and assess the performance of our approach with remotely acquired gait data [97] and wearable sensors [113]. Further, observed correlations of gait parameters with disability may help identify older PwMS advancing into sudden worsening, which may provide improved personalized care, and merits future investigation.

2.5 Summary

We present GML4MS, a novel machine learning pipeline for classification of PwMS using gait dynamics. The expression of MS over time and aging is heterogeneous, making the identifica- tion of sudden changes in PwMS, particularly difficult. In this work, we extracted normalized spatiotemporal and kinetic gait features and demonstrated the benefits of regress-N to differentiate MS and disability related changes. Further, we evaluated the effectiveness of GML4MS to generalize across different walking tasks and subjects. With a larger data set, generalization of subjects in one test environment to new subjects in a different environment would need to be validated. The current study on prediction and progression space in MS may aid neurologists to understand advancing disease with aging and identify meaningful ML-based strategies for identifying PwMS. Given that we have more older adults with MS than younger adults, and the expected continual shift of the peak prevalence of MS into older age groups, the prediction of a tipping point for older PwMS advancing into sudden worsening may provide improved personalized care. Early detection of these inflection points in older PwMS may lead to concise and effective detection strategies and in turn benefit both patients as well as clinicians to curtail MS therapy expenses.

CHAPTER 3

LEARNING A VISION-BASED FRAMEWORK

In this chapter, we review the work in A Vision-Based Framework for Predicting Multiple Sclerosis and Parkinson's Disease Gait Dysfunctions - A Deep Learning Approach

3.1 Introduction

Neurological gait disorders are associated with an increased risk of falls in older adults [7]. Abnormal gait has been observed in 35% of older adults, and associated with a greater risk of institutionalization and mortality [8]. While gait evaluation is common [11], few studies have focused on the differentiation of neurological disorders, such as Parkinson's disease (PD) or multiple sclerosis (MS), using gait analysis [118, 119]. Various gait evaluations, such as motion capture during the timed 25 foot walk and timed up and go test have been explored in clini- cal settings to assess neurological conditions, such as MS [120, 121] and PD [122]. Typically, specialized equipment like a lab-based motion capture system, force plate or electromyography sensors often is needed for these clinical quantitative gait measures, which can be expensive and require skilled personnel to analyze. Recent work on movement analysis with wearable iner- tial measurement unit sensors [123], smartwatches and smartphones [97] has overcome some of these constraints, yet these approaches are not contact-free and may require installation of multiple sensors. Past studies have explored depth cameras for gait monitoring [124], but these are relatively costly and not as easy to use. Herein, we used a standard RGB digital camera to examine pathological gait. This proposed system allows for passive and remote gait monitoring at reduced cost and effort, which should aid in making it a viable point-of-care technology for early detection of gait alterations in real-world settings. Moreover, we apply computer vision and deep learning (DL) algorithms to process our gait videos and extract significant information for an automated and objective quantification of neurological conditions. Given the inherently complex gait dynamics with little-known direct descriptors for the disorders, hand engineering

of features in this situation is complicated. DL automates this process of feature extraction and eliminates the need for domain expertise to allow for a remote real-time application, possibly at homes, of our entire workflow.

This study introduced and examined a vision-based gait analysis framework using DL for MS and PD gait dysfunction prediction. We extend prior work examining MS-related variations in spatiotemporal and kinetic gait characteristics [72, 91]. We classify strides of persons with MS (PwMS), healthy older adults (HOA), and persons with PD (PwPD) across three classification designs (see Figure 3.1):

1) Task generalization: We train ternary (HOA, PwMS or PwPD) classifiers on walking (W) trials (tasks) and use them to test on walking-while-talking (WT) trials. This experimental paradigm might be useful in quantifying how algorithms trained on normative data collected in a supervised lab or clinic could be used as a basis to assess gait data collected in a real-world home-based setting with challenges of divided attention.

2) Subject generalization: We train ternary classifiers on a balanced subset of subjects and use them to test on the remaining subjects. These algorithms may be useful in detection of disease in new patients.

3) Task-subject generalization: We train our classifiers on a balanced subset of subjects in W trials and use them to test on the remaining subjects in WT trials. This generalization framework is designed to simulate how algorithms could be used to predict disease in new subjects in more real-world settings.

3.1.1 Related Work

Past studies have quantified the decline of gait performance in PwMS [86]. Most gait-based approaches for MS detection have been based on statistical analyses of kinematic and kinetic data [87, 90]. Several recent works have applied traditional supervised machine learning (ML) to classify MS using gait data collected via treadmill [72], smartwatches and smartphones [97]. Vision methodologies based upon digital cameras have also been used to estimate clinical gait parameters in human gait analysis [125,126] and categorize other neurological populations [119, 127]. Depth cameras capturing 3D movement patterns have been explored for gait assessment in subjects with motion difficulties [128, 129], but those systems require a relatively costlier hardware, have some limitations when used outdoors and are constrained by the camera to object distance. Our contribution is using DL with a multi-view digital camera-based gait analysis framework for prediction of gait-related neurological disorders. Of particular novelty is our

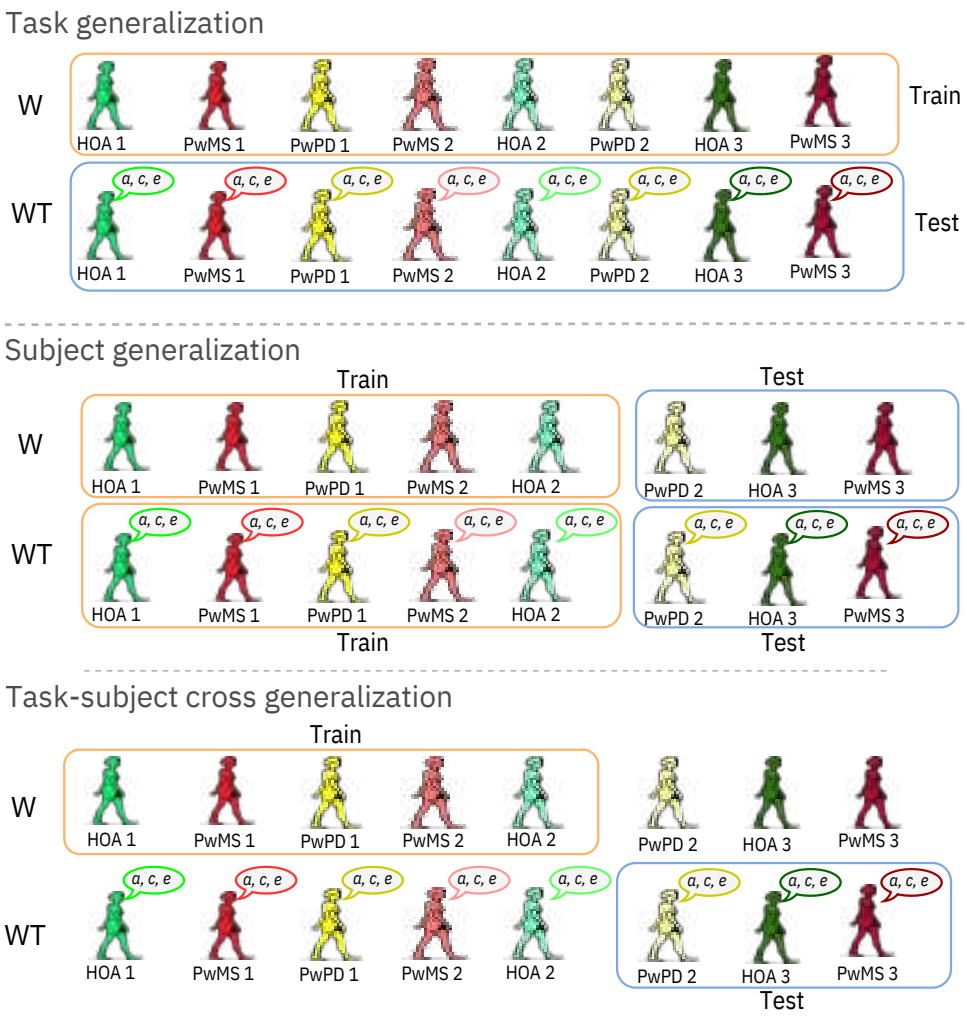

Figure 3.1: Top: Task generalization, Middle: Subject generalization, Bottom: Task-subject cross generalization. HOA, PwMS and PwPD are depicted in shades of green, red and yellow, respectively. The indices (1, 2, 3, . . .) are used as a reference for dummy subject identifiers.

focus on MS. We further considered a dual-task walking paradigm and consequently, a task-subject generalization classification framework. Most prior work has been focused on binary healthy-vs.-pathological gait [72, 97, 119]; we investigated a more challenging multi-class setup which further involves distinguishing between different causes of the neurological gait. Unlike past studies [119, 127], we have added feet features along with other body coordinates in our analysis.

The proposed application of vision and DL to learn gait dynamics in PwMS and PwPD across tasks, and over new subjects is a step towards the identification of worsening of symptoms in the near term. Our system requires only an inexpensive digital camera, and thus can be easily and economically deployed in homes of older adults for a real-time gait analysis with negligible

user interaction. We provide a comprehensive quantitative comparison of 16 diverse ML and DL algorithms for all classification designs (Figure 3.1); which may assist researchers in the selection of suitable model architectures and hyperparameters. Moreover, we discussed the global and local importance of our extracted features in the classification performance; and explored a potential association between our model predictions and the lower extremity function of subjects.

The remainder of the chapter is organized as follows. We start out in Section 3.2.1 by introducing the data acquisition paradigm. In Sections 3.2.2-3.2.4, we discuss our data analysis methodology, comprising of video processing, feature designs and classification strategies. In Section 3.3, we present our model prediction results and post hoc analysis. Finally in Sections 3.4 and 3.5, we highlight some concluding remarks along with limitations and future directions for this study.

3.2 Methods

3.2.1 Experimental Design: Subjects and Setup

3.2.1.1 Studyparticipants

The study consisted of 33 participants: 10 PwMS (age: 66 ± 9 years, 6 male), 9 PwPD (age: 68 ± 5 years, 3 male), and 14 HOA (age: 63 ± 9 years, 3 male). All participants were medically stable, had a cognitive status score [106] of above 18, were right-hand dominant, had no lower limb injury in the past six months, and had normal or corrected-to-normal vision. PwMS had mild to moderate disability as evaluated by the Kurtzke Expanded Disability Status Scale [107] [2.0 – 6.0], were relapse-free for 30 days prior to experimental trials and had no other cognitive dysfunction or neurological disorders. PwPD had mild to moderate severity on the Hoehn & Yahr Scale [130] [1.0 – 4.0], were "ON" anti-Parkinsonian medication state and had no other cognitive dysfunction. See Table B.1 in the Appendix for additional details on subject demographics. Prior to testing, all participants provided informed consent approved by the local Institutional Review Board (Protocol No. 15674).

3.2.1.2 Experimentalparadigm All subjects performed two self-paced walking tasks on an instrumented treadmill (C-Mill, Motekforce Link): 1) a trial in single-task walking (W) and 2) a trial in dual-task walking-while-talking (WT) condition. For the WT task, subjects were asked to walk and recite alternate letters

of the alphabet while providing an equal priority to both walking and talking. Two 800 ×448 pixels resolution digital cameras were located facing subject's front and right side to record their lower half and feet movements (resp.) at 30 frames per second. Given prior evidence of increased variability in footfall placement in PwMS [131], we focused our cameras on subject's feet and lower extremity in this study (Figure 3.2). All extracted gait videos were truncated to 60 seconds, to account for alterations in gait speed during gait initialization and gait arrest. Additionally for validation, CueFors 2 software was used to collect gait events and raw center of pressure position coordinates at 500 Hz during each walking trial. A total of 116 gait videos, combining subject's front- and side-views, were gathered for 33 (W: 32, WT: 26) subjects.

3.2.2 Data Analysis: Gait Video Processing

The proposed data analysis pipeline for this work is presented in Figure 3.2.

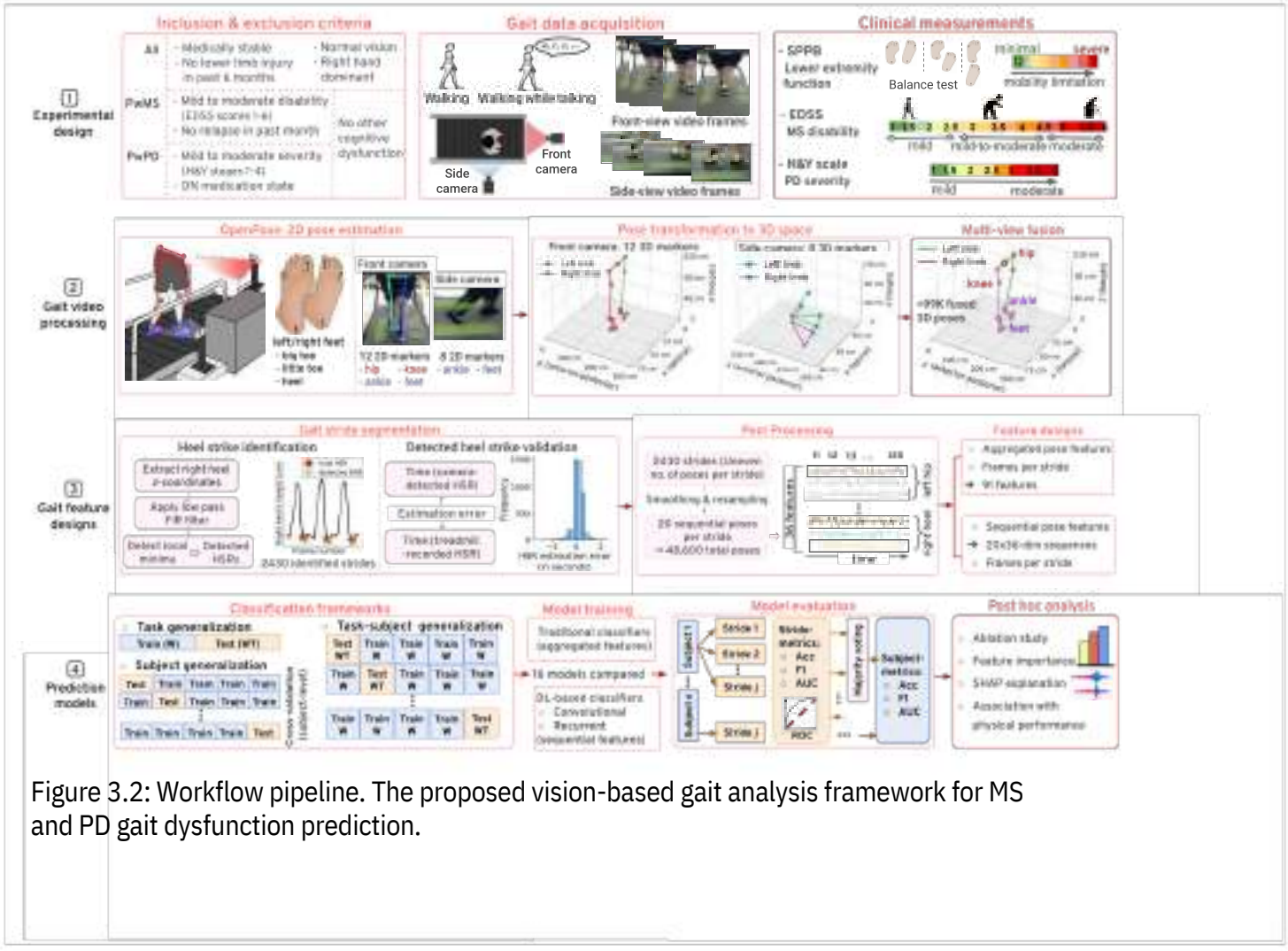

Figure 3.2: Workflow pipeline. The proposed vision-based gait analysis framework for MS and PD gait dysfunction prediction.

3.2.2.1 2Dbodyposeestimation OpenPose [132] was used to locate the 2D pixel coordinates, estimating the skeletal joint positions of a detected subject in each frame of the collected gait videos. By connecting the anatomical key points within frames, we obtained a consecutive sequence of identified biomechanical poses capturing the subject's unique gait patterns along each walking trial. OpenPose provides an open-source real-time architecture for robust body pose estimation [133] using a fine-tuned VGG-19 convolutional neural network (CNN) [134]. VGG-19 CNN produces a set of feature maps F for input to the first stage of a multi-stage CNN architecture. The iterative multi-stage CNN framework predicts a set S of 2D confidence maps of joint locations and a set L of 2D vector fields of part affinity fields (PAFs) that encode the degree of association between different joints, while concatenating predictions from the prior stage and original features F to improvise predictions over subsequent stages. Specifically, the first set of stages predicts PAFs, while the last set predicts confidence maps, i.e., mathematically, $L1 = \psi 1(F)$, $Lt = \psi t(F, Lt-1) \ \forall \ 2 \le t \le TP$, where ψt and TP denote the CNNs at stage t and total number of PAF stages, respectively; and after TP iterations, $S^{TP} = \phi(F, L^{TP})$, $S^{t} = \phi t(F, LTP, St-1) \ \forall \ TP < t \le TP+TC$, where ϕt and TC denote the CNNs at stage t and total number of confidence map stages, respectively. A mean squared error (L2 loss), between the estimated predictions and the ground truth maps and fields, is applied at the end of each stage. Lastly, a greedy inference is applied over the confidence maps and PAFs to yield the subject's 2D joint key points. OpenPose generated the 2D location coordinates and corresponding prediction confidence of 12 front-view lower extremity landmarks (i.e., hips, knees, ankles and foot keypoints) and 8 side-view ankle and foot landmarks for both sides of the body (see Figure 3.2). OpenPose may occasionally generate erroneous poses with left and right sides swapped, missing keypoints, or falsely perceived human body due to a range of possible reasons, including self-occlusion, varying lighting or color information. Thus post-processing involved correcting for switches between the left and right limbs, quadratic interpolation of missing markers, and identification of erroneous landmarks. Following processing, 2D skeletal landmark coordinates (in pixels) were retained from 102,598 front- and side-view gait postures.

3.2.2.2 Posetransformationto3Dglobalcoordinatespace Camera calibration was carried out, using intrinsic and extrinsic camera matrices, to transform the estimated front- and side-view 2D joint locations in local image pixel coordinates to 3D positionsinaglobalcoordinatesystem,denotedby $(xw^{(k,i)}, yw^{(k,i)}, zw^{(k,i)})$ forkeypointkincamera view i. To computationally approximate intrinsic and extrinsic camera parameters, we calibrated

both of our cameras using sample patterns from 3D real world position and corresponding 2D image coordinates of square corners in a chess board. Post-processing consisted of bounding all computed 3D position coordinates, using real-world constraints (i.e., treadmill width and length, and height of person). An example of computed front- (red markers) and side-view (blue markers) 3D poses is shown in Figure 3.2.

3.2.2.3 Multi view fusion of 3D body poses We conducted a weighted mean-based multi-view fusion, as proposed in [127], of 3D joint positions across views in the two planes; this helps to account for deviations in 2D pose approximation (3.2.2.1) and 3D transformation (3.2.2.2). Essentially, our weights are normalized prediction probabilities (as estimated by OpenPose) of detecting a joint at the respective coordinates, i.e. we assign higher weight to the view that estimates 2D coordinates for the joint with greater confidence. Mathematically, let p(k,i) denote the OpenPose estimated confidence for key point k in view i, then for every key point k common in both views, namely, left and right ankles, big toes, little toes and heels, we compute the multi-view fused 3D coordinates as:

$$[x(k)_w, y_w^{(k)}, z_w^{(k)}]^T = \sum_{i \in \{front, side\}} w(k,i)\,[x(k)_i, y_{w}^{(k,i)}, z_{w}^{(k,i)}]^T,$$

$$\text{where weights } w(k,i) = \frac{p(k,i)}{\sum_{i \in \{front,\, side\}} p^{(k,i)}}$$

Only frames with both front- and side-view pose available were merged. Subsequently, 36 (3 × 12 joints) body keypoint features were derived from a total of 99,942 multi-view (x, y, z) fused poses, split into 57,708 (HOA: 28,174, PwMS: 16,210, PwPD: 13,324) and 42,234 (HOA: 13,763, PwMS: 13,572, PwPD: 14,899) poses across 32 and 26 subjects in trials W and WT (resp.). See Figure 3.2 for an illustration of a fused 3D pose.

The demographic variations between subjects, especially their height, may inherently bias their gait dynamics and so influence the neurological gait differentiation ability. Therefore, to standardize for subject heights in our data, we scale pose coordinates to normalize the average of estimated left and right hip heights to a constant, particularly 100 cm, which is about American median hip height [135]. In addition to primary checks on bounds of coordinates ($0 \le x_w^{(k)} \le 87$, $0 \le y_w^{(k)} \le 310$ and $0 \le z_w^{(k)} \le h$), we further did sanity tests on distributions of extracted features to ensure they align with usual human body dimensions.

3.2.2.4 Validatingestimated3Dposesthroughtreadmill'scenterofpressure

• Validation procedure: Using treadmill center of pressure data and detected gait events using force plate data as a ground truth during single support and dual support stance in a gait cycle, the average center of pressure position was calculated. Using this estimate, the difference with the average centroid of the estimated base of support during single and dual support stance in a gait cycle were calculated. Let's set up some notation. A stride is characterized by the following gait events in order: HSR: heel strike right, TOL: toe-off left, MidSSR: midstance right, HSL: heel strike left, TOR: toe-off right, MidSSL: midstance left, with the next HSR starting a new stride. For each subject s, event e of stride n occurs at time $t_{e,n}^{s}$; we have

$$t_n^{HSR,s} < t_n^{TOL,s} < t_n^{MidSSR,s} < t_n^{HSL,s} < t_n^{TOR,s} < t_n^{MidSSL,s} \quad (3.1)$$

for $n \in \{1, 2 \ldots N_s\}$, where N_s is the number of strides of subject s. We also have that $t_n^{MidSSL,s} < t_{n+1}^{HSR,s}$ for $n \in \{1,2 \ldots N_s - 1\}$. By shifting, we assume that $t_1^{HSR,s} = 0$, which corresponds to the first HSR after both treadmill and video camera are actively recording data. Ground truth stride events are defined here as detected by the treadmill. The treadmill data is sampled every $\delta_T \stackrel{def}{=} 0.002$ seconds. Fixing a subject s, the treadmill logs a center $\bar{p}^{(s)}(k\delta_T) = (\bar{x}^{(s)}(k\delta T), \bar{y}^{(s)}(k\delta T)) \in R^2$ of pressure (with the first coordinate denoting lateral position and the second denoting anterior-posterior position) at time $k\delta T$. We assume (via centering) that $\sum_k p(k\delta T) = 0$ for each subject s; this compensates for positional bias on the treadmill. The point $\bar{p}^{(s)}(k\delta T)$ is the center of gravity of the forces measured by the treadmill force plate at that time. Video data is sampled every $\delta_V \stackrel{def}{=} 1/30$ seconds (a lower sampling rate than the treadmill). Fixing a subject s, we have an estimated position vector (from 3.2.2.3)

$$p^{(s)}(k\delta V) = \big(\big(x_{L,lt}^{(s)}(k\delta V), y_{L,lt}^{(s)}(k\delta V)\big), \big(x_{L,bt}^{(s)}(k\delta V), y_{L,bt}^{(s)}(k\delta V)\big), \big(x_{L,h}^{(s)}(k\delta V), y_{L,h}^{(s)}(k\delta V)\big), \big(x_{R,lt}^{(s)}(k\delta V), y_{R,lt}^{(s)}(k\delta V)\big), \big(x_{L,bt}^{(s)}(k\delta V), y_{R,bt}^{(s)}(k\delta V)\big), \big(x_{R,h}^{(s)}(k\delta V), y_{R,h}^{(s)}(k\delta V)\big)\big) \in (R^2)^6$$

of estimated lateral and anterior-posterior positions of the little toe (lt), big toe (bt), and heel (h) for left (L) and right (R) feet. As with treadmill data, we assume (via centering)

that

$$\sum_{\substack{k\in\{lt,bt,h\}\\ f\in\{L,R\}}} \left(s_{f,c}[k\delta] - \left(x_{f,c}^{(s)}(k\delta), y_{f,c}^{(s)}(k\delta)\right)\right) = 0.$$

to again remove positional bias. If our estimated pose were correct, the treadmill reading $\bar{p}^{-}(s)$ would be the force-weighted average of the components of $p(s)$ which are in contact with the treadmill at that moment. We can divide time for each subject into sets where the subject has single support (left and right) or double support, depending on whether one or both feet are in touch with the treadmill force plate; namely, for subject s, we have the sets

$$\begin{aligned} SSR(s) &= \bigcup_{n=1}^{N_s-1} \{(t_n^{TOL,s}, t_n^{HSL,s})\} \\ SSL(s) &= \bigcup_{n=1}^{N_s-1} t_n^{TOR,s}, t_{n+1}^{HSR,s} \\ DS(s) &= \{(t_n^{HSL,s}, t_n^{TOR,s})\}_{n=1}^{N_s-1} \cup \{(t_n^{HSR,s}, t_n^{TOL,s})\}_{n=1}^{N_s-1} \end{aligned} \tag{3.2}$$

of stride intervals. See Figure 3.3 for an illustration of our validation procedure. In the single-support sets, when only one foot is on the force plate, the treadmill averages forces over three points (lt, bt, and h of the relevant foot). In the double-support sets, when both feet are on the force plate, the treadmill averages forces over all six points (lt, bt, (and hof both f)eet simultaneously). Within each stride interval I for each subject, e.g., $t_n^{TOL,s}, t_n^{HSL,s}$, we average $\bar{p}^{-}(s)$ and $p(s)$, extending the definition of $\bar{p}^{-}(s)$ and $p(s)$, define

$$\begin{aligned} \bar{p}^{-}(s)(I) &\stackrel{def}{=} \sum_{k:k\delta T\in I} \bar{p}^{-}(s)(k\delta T) \Big/ |\{k : k\delta T \in I\}| \\ p(s)(I) &\stackrel{def}{=} \sum_{k:k\delta V\in I} p(s)(k\delta V) \Big/ |\{k : k\delta V \in I\}| \end{aligned} \tag{3.3}$$

(using the standard interpretation of R2 and (R2)3 as vector spaces). Let's now compare $\bar{p}^{-}(s)(I)$ to the appropriate centroid of the base of support, a convex hull whose vertices are the given points of contact in $p(s)(I)$. Specifically, define

$$\left(\check{x}^{(s)}(I), \check{y}^{(s)}(I)\right) \stackrel{def}{=} \begin{cases} \frac{1}{l}\left(x_{L,c}^{(s)}(I), y_{L,c}^{(s)}(I(I))\right) & \text{if } I \in SSL(s) \\ \frac{1}{3}\sum_{c\in\{t,bt,h\}}\left(x_{R,c}^{(s)}(I), y_{R,c}^{(s)}(I)\right) & \text{if } I \in SSR(s) \end{cases}$$

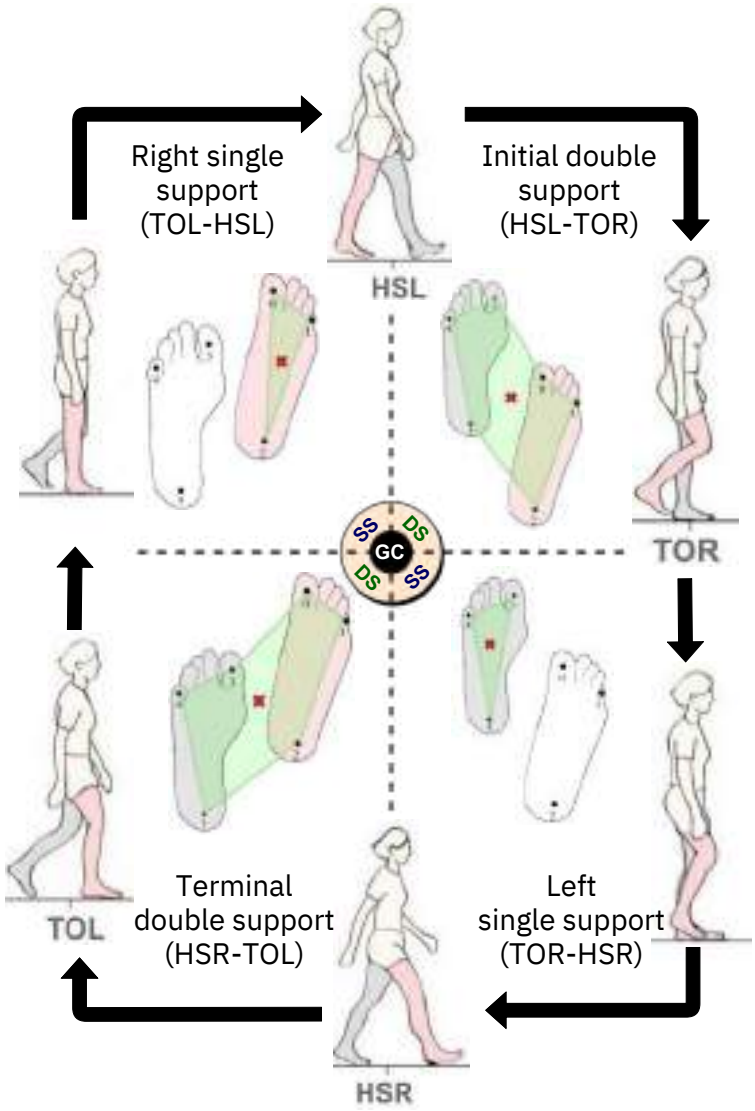

Figure 3.3: Base of Support and it's centroid throughout a stride starting at HSL. Subject's right (pink) and left (grey) feet keypoints, namely, lt, bt, and h, are labelled (in order) as 0, 1, 2 and 3, 4, 5, respectively. The overground view of feet is shown for each support phase, where foot not in contact with ground is colored white. The green shaded regions, i.e. hexagon for the double support and triangle for the single support, define the base of support and red × mark is it's centroid. Abbreviations: HSR: heel strike right; TOL: toe-off left; HSL: heel strike left; TOR: toe-off right; SS: single support; DS: double support.

and define

$$\check{x}^{(s)}(I) = \{6 \quad x^{(s)}_{f,c_i}I(\quad) + x^{(s)}_{fi+1,ci+1}(I)) \times (i=1 \; (Iy)\; {}^{(s)}_{fi+1,ci+1}(I) - x^{(s)}_{fi+1,ci+1}(I)\, y^{(s)}_{fi,ci}(I))\} / x f \sum i,c6(\; {}^{(s)}_{3i=1} \; x^{(s)}_{fi,c(}I)\, y^{(s)}_{fi+1,ci+1}(I) - x^{(s)}_{fi+1,ci+1}(I)\, y^{(s)}_{fi,ci}(I)\,)\},$$

$$\check{y}^{(s)}(I) \overset{\text{def}}{=} \{\sum_{i=1}^{6} (y^{(s)}_{f_i,c_i}(I) + y^{(s)}_{f_{i+1},c_{i+1}}(I)) \times (x^{(s)}_{f_{i+1},c_{i+1}}(I) - x^{(s)}_{f_i,c_i}(I))\} / (\sum_{i=1}^{6} (x^{(s)}_{f_i,c_i}(I) + x^{(s)}_{f_{i+1},c_{i+1}}(I)) \times (x^{(s)}_{f_{i+1},c_{i+1}}(I) - x^{(s)}_{f_i,c_i}(I)))$$

with the x and y values on the right-hand side corresponding to the components of points of $p^{(s)}(I)$ in (3.3), and where $(f_i, c_i)_{i=1}^{7} = ((L, lt), (L, bt), (L, h), (R, lt), (R, bt), (R, h), (L, lt))$. Expanding $\bar{p}^{(s)}(I)$ in (3.3) as $(\bar{x}^{(s)}(I), \bar{y}^{(s)}(I))$, we define lateral, anterior-posterior, and Euclidean distances

$$\begin{aligned} d^{(s)}_{L}(I) &\overset{\text{def}}{=} \check{x}^{(s)}(I) - \bar{x}^{(s)}(I) \\ d^{(s)}_{AP}(I) &\overset{\text{def}}{=} \check{y}^{(s)}(I) - \bar{y}^{(s)}(I) \\ d^{(s)}_{E}(I) &\overset{\text{def}}{=} \sqrt{(d^{(s)}_{L}(I))^2 + (d^{(s)}_{AP}(I))^2} \end{aligned} \tag{3.4}$$

which measure how well the pose-estimation agrees with treadmill data. The d(s)'s encompass several errors:

– discrepancy between open pose estimates of foot positions and ground truth points of contact

– unequal distribution of weight between the points of contact

– temporal fluctuations in weight.

We can now compute means of the distances of (3.4) in various ways. In doing so, we consider only intervals which have endpoints in the proper order of (3.1) (including $t^{MidSSL,s}_{n} < t^{HSR,s}_{n+1}$ as appropriate); if the treadmill fails to detect a stride event or the stride events are out of order, we pass to the next valid interval. Overall, 2483 strides with 9768 (single support: 4802, double support: 4966) valid support phases were retrieved across all synced subject videos in both W and WT trials and used to validate the centroid of our estimated base of support against the treadmill's center of pressure.

• Validation results: Quantitatively, for single support samples, the aggregated (over all videos) mean and standard deviation of Euclidean, lateral and anterior-posterior distances (in cm) were 9.95 ± 5.68, 0.04 ± 4.85 and 0.61 ± 8.96 (resp.); and similarly, 8.82 ± 5.49, −0.04±3.10 and −0.57±8.53 (resp.), for the dual support samples. While the center of pressure of the participant and centroid of the base of support are not expected to be

perfectly aligned, we found congruence between these measures, which helped reaffirm the validity of our estimated 3D poses.

3.2.3 Data Analysis: Gait Feature Designs

For our ML and DL classifiers, we derived features across individual strides. This stride-wise feature extraction approach extracts multiple samples from a single subject; thus augmenting and introducing significant variations to our dataset to improve the generality of ML and DL learners. Moreover, stride-wise predictions allow for frequent and even near real-time inferences for potential clinical applications.

3.2.3.1 Gaitstridesegmentation After fusing 3D poses (3.2.2.3), we performed automatic gait stride segmentation. In order to do
so, we detected heel strikes on the right side of body (HSRs) that conventionally mark the start of everynewstride.Let $^{(k)}$ $^{(k)}$ $[xw,yw,zw]^{(k)}$denotethefused3Djointpositioncoordinatesforkeypoint k. Then, we defined HSRs as the local minimas (at least one second apart) in the filtered right heelheightseries, $zw^{(rightheel)}$.

• Heel strike detection: Overall, 2430 strides were retrieved from 33 (HOA: 14, PwMS: 10, PwPD: 9) subjects across three cohorts and two trials. More specifically, 1380 (HOA: 658, PwMS: 389, PwPD: 333) and 1050 (HOA: 351, PwMS: 332, PwPD: 367) strides were retrieved from 32 and 26 subjects in trials W and WT (resp.). HOA, PwMS and PwPD had on average 47.0 ± 7.9, 38.9 ± 8.3, 41.6 ± 2.1 strides and 43.9 ± 2.8, 36.9 ± 9.6, 40.8 ± 3.9 strides in trials W and WT (resp.). Subjects from the same cohort on average walked fewer strides in the more challenging WT task than in the W task. Healthy subjects had more strides than impaired in both the trials. Next, we discarded all the poses before the start of the first stride and after the end of the last stride. Thus out of 99,942 (W: 57,708, WT: 42,234) multi-view fused poses (in Section 3.2.2.3), now, 56,226 (HOA: 26,541, PwMS: 16,187, PwPD: 13,498) and 41,747 (HOA: 13,638, PwMS: 13,448, PwPD: 14,661) poses remained across 1380 and 1050 detected strides in trials W and WT (resp.). HOA, PwMS and PwPD averaged 40.3 ± 10.0, 41.6 ± 9.4, 40.5 ± 8.8 frames and 38.9 ± 9.2, 40.5 ± 8.5, 39.9 ± 8.6 frames per stride in trials W and WT (resp.). A higher frame count per stride indicates a slower stride speed, i.e., HOA on average walked with an increased gait speed in both trials.

• Heel strike validation: To quantify the performance of our HSR detection procedure, we begin with using the treadmill-recorded gait event data to mark frames with true HSRs. This required syncing video and treadmill records for each subject-trial pair (for details on the sync process, refer 3.2.2.4). Heel strike identification segment in Figure 3.2 plots a snippet of the filtered right heel height series for a PwMS with true and algorithmically detected HSRs shown in red stars and green diamonds (resp.). We define the HSR estimation error as the time gap (in seconds) between the pose-estimated HSR and the corresponding closest true HSR. The error is positive for a late and negative for an early estimate of the HSR. Heel strike validation segment in Figure 3.2 depicts the frequency distribution of estimation errors across all our subjects. Overall, detected HSRs were on average 0.125 ± 0.35 seconds late relative to ones recorded via treadmill. In general, a good correspondence was attained between true and identified HSR markers across HOA as well as PwMS and PwPD with likely gait irregularities.

• Heel strike normalization: Following stride segmentation, we had a varying number of poses grouped by stride. Thus we carried out temporal down sampling and smoothing (using a disjoint window-based moving average approach) in order to normalize poses to 20 per stride. Ultimately, we had 1380 (HOA: 658, PwMS: 389, PwPD: 333) and 1050 (HOA: 351, PwMS: 332, PwPD: 367) strides (data samples) in trials W and WT (resp.), where each stride was a 20× 36-dimensional sequence with 36 body keypoint coordinates (features/time series) over 20 consecutive time-normalized frames (time steps).

3.2.3.2 Featuredesigns

• Aggregated pose features: We utilized descriptive statistical measures to aggregate our 2D (20 × 36) strides along the time dimension into a vector with 91 features. These aggregated features were used with traditional ML classifiers (logistic regression, random forest, etc.) that expect a flattened feature vector as input data. In particular, to assess deviation in a stride, we compute the coefficient of variation and range for 36 joint coordinate series, each with 20 time steps; hence, obtaining 72 aggregate features. Further, to estimate mismatch in gait between left and right sides of the body, we measured asymmetry as absolute difference in the range of left and corresponding right keypoint coordinate series; thus securing 18 (3 (x,y,z) × 6 joints) more features. Finally, we included the original number of frames per stride (before resampling to 20 poses in 3.2.3.1) as a feature indicative of subject's gait speed; thereby totaling to 91 variation-, asymmetry- and speed-based characteristics in each stride to distinguish gait variations in controls from

neurological population. As a result, we gathered a dataset with 2430 strides (data samples/rows) across W and WT and 91 features (columns) to feed into our traditional ML classifiers.

• Sequential pose features: In contrast to traditional ML models, DL-based classifiers do take 2D sequential keypoints data straight away as input; therefore, we did not carry out any additional feature engineering for our strides. This configuration did not risk losing information during the aggregation of features. Given the temporal fluctuations and irregularities in gait features within a stride, DL classifiers should be able to leverage this sequential information to generate improved predictions. Similar to the aggregated pose features, we included the original number of frames per stride as an additional feature, demonstrative of gait speed, to the model's input. This resulted in 2430 strides (data samples) across W and WT, each consisting of a 36-dimensional sequence over 20 consecutive time steps and scalar speed, as input for the DL algorithms.

3.2.4 Data Analysis: Classification and Evaluation

We utilized the designed features to classify unique gait dynamics in HOA, PwMS and PwPD on a stride-by-stride basis. We used nine traditional supervised ML algorithms to establish baseline performance: logistic regression (LR), support vector machine with linear (LSVM) and radial basis function (RBF SVM) kernels, decision tree (DT), random forest (RF), adaptive boost- ing (AdaBoost), eXtreme gradient boosting (XGBoost), gradient boosting machine (GBM) and multilayer perceptron (MLP). All these classifiers required 1D feature vector and thereby the aggregated pose features (3.2.3.2) are used as their input. Z-score normalization was applied to all aggregated features to eliminate the influence of variable feature ranges in the model's input.

3.2.4.1 Deeplearningclassifiers:Convolutionalarchitectures

In this segment, we describe the 4 convolutional DL models used in our study. For these algorithms, temporal data with concatenated features over 20 consecutive frames (3.2.3.2) was presented directly as input. We used Z-score normalization before feeding in data to the models.

• 1D convolutional neural network (CNN): Our 1D CNN model included b convolutional blocks (ConvBlocks), where b is a tuned hyperparameter; each ConvBlock consisted of a 1D convolutional layer (ConvLayer) followed by batch normalization, non-linear activation, dropout [136] and a pooling operation. A 1D ConvLayer with f 1D convolving filters

(trainable parameters), each with a filter length of k (kernel size: k × 36) and convolution stride s (steps by which the filter moves) maps our 20 × foutputsequence,wherefandsarehyperparameters. × 36 input sequence in the first con-

vBlock to a new $(\frac{20-k}{s}+1)$

Let X_{b-1} be the output for the (b − 1)-th ConvBlock, W b, f and bias b be the weight matrix of the f-th filter and bias vector (resp.), for the b-th ConvLayer, then, the corresponding output for the b-th ConvLayer is given by:

$$H_{b,f}(t) = \sum_{c=0}^{Nc} \sum_{i=-k}^{k} X_{b-1}(t - i, c) W_{b,f}(i,c) + bias_b(t),$$

where Nc is the number of features in the (b − 1)-th ConvLayer, i.e. Nc = 36 for the first ConvLayer and kernel length k is expressed as 2k + 1. ConvLayers take advantage of sparse connectivity and further impose local connectivity within proximate neural units, to lessen the parameters learnt as well as the chances of overfitting. In essence, the convolution function hierarchically extracts low-level features from the input data in the initial few ConvLayers to more complex high-level characteristics as the number of layers advance. We used batch normalization to standardize the input for the subsequent ConvLayer over each batch in the training process; it helps expedite training by offering some regularization.Mathematically,letBbethemini-batchwith m trainingsamples, $\mu_B(t)$ and $\sigma^2_B(t)$ be thecorrespondingmeanandvariancea t time step t, then, $\tilde{H}_{i\,b,f}(t)\ \forall\ 1 \le i \le m$denotes theper-dimensionnormalizedoutputofthe-thConvLayer:

$$\mu_B(t) = \frac{\sum_{i=1}^{m} H_{i\,b,f}(t)}{m} \qquad \sigma^2_B(t) = \frac{\sum_{i=1}^{m} (H_{i\,b,f}(t) - \mu_B(t))^2}{m};$$

$$\tilde{H}_{i\,b,f}(t) = \frac{H_{i\,b,f}(t) - \mu_B(t)}{\sqrt{\sigma^2_B(t) + \epsilon}}, \qquad 1 \le i \le m$$

where ϵ is an arbitrarily small constant added for numerical stability. Following normalization, we applied an activation function to introduce non-linearity into ConvLayer's out- put neurons. Though we experimented with several non-linearities with diverse merits and limitations, a rectified linear unit (ReLU), computed as ReLU(x) = max(0, x), is amongst the most frequently used activation, for it does not saturate or cause vanishing gradients. Further, dropout disables neurons and their corresponding connections at ran- dom in the model with probability p (hyperparameter) to help prevent overfitting during training. Additionally, a pooling (sub-sampling) layer is intermittently included in between ConvLayers to lower the number of model parameters, and thus, manage overfitting; max pooling preserves maximum value from a bunch of r neurons, thus dividing the current

dimensionality by r. This also ensures ConvLayer-extracted features are invariant to minor translations in the input data. Following these b ConvBlocks, the 2D output ($l \times h$) is flattened to a vector of length either lh or h (via global average pooling, where we only retain the average of each feature map). The additional frames per stride feature is now concatenated with the 1D model output vector and passed through multiple DenseBlocks. Each DenseBlock consisted of a fully connected layer with a non-linear activation at the outcome, except for the last layer. Note that our final linear layer yields a vector of length 3, same as the number of categories.

Since CNNs do not include any recurrence mechanism, we used positional encoding to explicitly add information with each pose about its corresponding order in the input stride.

Specifically, let $x_t \in R^{36}$ be the feature vector for the t-th pose ($0 \le t < 20$) in the stride and $p_t \in R^{36}$ be the corresponding positional encoding vector, then, $x'_t = x_t + p_t \; \forall \; 0 \le t < 20$ is the upgraded embedding that is fed as input to the model. We used the sinusoidal encoding [137] that generated pt as follows:

$$pt(j) \overset{def}{=} \begin{cases} \sin(t/10000^{2k/36}), & \text{if } j = 2k \\ \cos(t/10000^{2k/36}), & \text{if } j = 2k+1 \end{cases},$$

where $t \in [0, 20)$ and $j \in [0, 36)$ denote the corresponding time step and index of the feature dimension (resp.). For sinusoidal positional encoding, the distance between adjacent time increments is symmetrical and declines well over time. It also enables to smoothly learn relative positions as $pt(j + \delta)$ can be expressed as a linear function of $pt(j)$ for any fixed δ. Further, it contributes negligible overhead to our computational expenses. as it merely adds sine and cosine functions to the input embedding. Moreover, employing positional encoding as an alternative to recurrence (with strong sequential dependencies) significantly expedites our training time.

- Residual neural network (ResNet): To extract more intricate features, deeper CNN networks are generally desired. However, deeper networks are increasingly challenging to train due to the degradation issue wherein as the model depth increases, its corresponding performance saturates and then degrades swiftly owing to a higher training error than its shallower counterpart. In theory, the deeper layers could simply be identity maps stacked to the corresponding shallow architecture to maintain just the same training error and thus, avoid any degradation in accuracy with added layers. ResNets precisely leverage this understanding and lets network layers explicitly learn residual functions relative to the layer inputs and thereby assist in the training of deeper models [138]. Let g(x) be the expected

function to be fit by a given stack of layers, where x indicates the input to the first layer. The residual connection learns def – x and later recasts the learnt mapping as f(x) + x via element-wise addition to recover the original function g(x). In case the dimensions of x and f(x) are unequal, we use a 1 × 1 ConvLayer to compute a linear projection W so as to resize x and compute f(x) + Wx. ResNets benefit from increased model depths by easing optimization and adding no extra parameter or computational cost. We experimented with two kinds of residual blocks, namely, basic and bottleneck blocks. A basic block is a stack of 2 1D ConvLayers, each followed with a batch normalization and a ReLU activation. Note that the second non-linearity was applied after the element-wise addition of the input with the learnt residual mapping. A 1 × 1 convolution is used on the input when required to match dimensions for the element-wise addition. The deeper bottleneck block is similar in design but with a stack of 3 ConvLayers instead of 2. Note that the number of filters f , corresponding filter length k and stride s are tuned hyperparameters for each ConvLayer. The 20 × 36 model input is first parsed using an initial ConvBlock, comprising a ConvLayer followed by batch normalization, ReLU activation and max pooling (in order), to embed features prior to residual blocks. Next, we used a stack of b (hyperparameter) basic or bottleneck blocks to set up residual learning within every few layers for deeper network designs. Eventually, the 2D output is flattened via global average pooling, concatenated with frames per stride and transformed using multiple DenseBlocks to a length 3 vector. We also experimented with using positional encoding with ResNets to augment order information to our input. Figure 3.4 shows a sample ResNet architecture (top right) along with the design for basic (top left) and bottleneck (bottom left) residual blocks.

- Multi-scale residual neural network (MSResNet): Given varying and noisy nature of extracted joint position coordinates amid our walking subjects, utilizing a fixed single-scale convolutional kernel size to extract features from only one scale may not be optimal. Consequently, We applied multi-scale kernel-based ResNet architecture, as proposed in [139], to derive deep-hierarchical features from multiple scales out of raw poses. MSResNet incorporates both the residual learning framework and multi-scaled convolutional kernels to address performance degradation issues and learn robust characteristics in multi-scale views from pose locations. Similar to ResNet, the input pose positions are firstly passed via a ConvLayer followed by batch normalization and ReLU activation. Next, we traversed the extracted features through three branches, each applying a different scale of convolutional kernels to acquire attributes from multiple receptive fields. Each branch is a stack of 3 basic blocks with {64, 128, 256} filters (resp.); filter lengths for

ConvLayers in three different branches were set to be 3, 5, and 7 (resp.). The batch of residual blocks in all branches is followed by a global average pooling layer to reshape output features into a flattened vector. The vectors from the three branches are then concatenated into a single vector of length 768 (= 256 × 3) and appended with the additional frames per stride feature; finally, this concatenated vector is fed to a fully connected network with 3 output units. Figure 3.4 depicts our MSResNet architecture (bottom right).

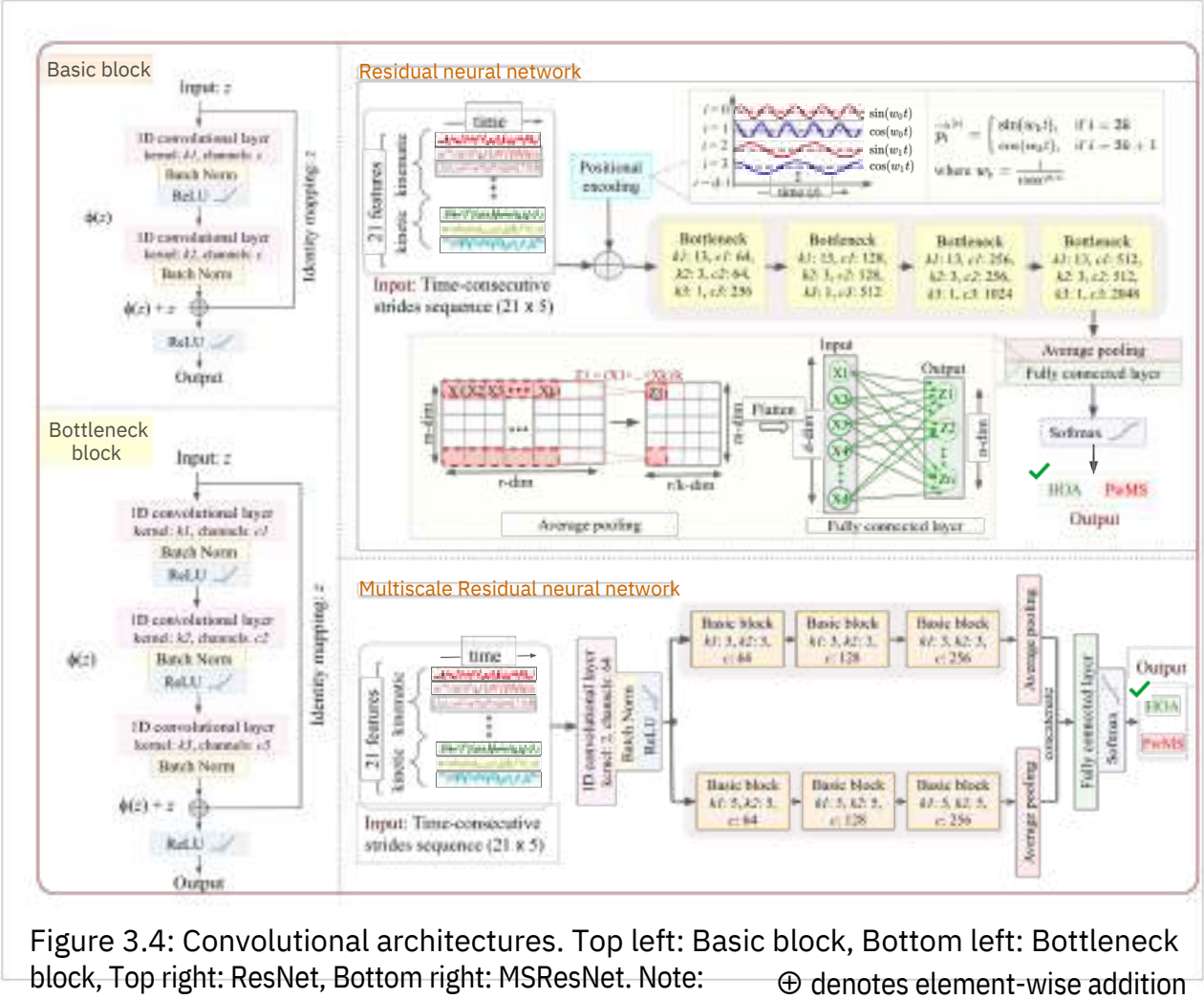

Figure 3.4: Convolutional architectures. Top left: Basic block, Bottom left: Bottleneck block, Top right: ResNet, Bottom right: MSResNet. Note: ⊕ denotes element-wise addition in the basic and bottleneck residual blocks.

• Temporalconvolutionalnetwork(TCN):RecentlyintroducedTCNs[140]havematched and even exceeded the performance of several recurrent models over numerous sequential modelling tasks. In general, the TCN architecture is relatively simpler, possesses longer memory to capture a more extended history and in practice, demands minimal tuning. TCNs have been lately used for time series assessment across multiple domains [141,142].

TCN employs 1) dilated causal convolutions to process temporal data, where causality ensures no data from the future is leaked to the past and dilations assist the network to form long histories through large receptive fields, , that enable to look quite far back in the past while making predictions, and 2) residual connections to train deeper models well. Our TCN model consisted of a stack of n (hyperparameter) TCN residual blocks. Each block first learnt the residual via 2 1D dilated causal fully convolutional layers, each followed by weight normalization [143], ReLU activation and a dropout layer (in order), and then, is further succeeded by another ReLU after the element-wise addition of the input with the estimated residual mapping. An additional 1 × 1 convolution is used on the input to ensure dimensions match for the element-wise addition. A fully convolutional layer is simply a ConvLayer with an output of the same size as the input; causal convolutions ensured that output at time t is convolved solely with features prior to and at time t in the earlier layer. Further, a convolution with dilation factor d oflengthkiscomputedas(g ∗ dfx)()= ∑onanelementxofa1Dinputgwithfilterf stepdbetweeneverytwoadjacentfiltertaps.Inpractice,wesetd=2iforthei-thlevel (TCN block) of our network; this allows the receptive field size to exponentially increase relative to the depth of the network. A convolution stride of 1 and zero padding of length (k − 1)d is added to the ConvLayer with kernel length k (hyperparameter) and dilation factor d to ensure fully convolutional behaviour. Next, we applied weight normalization to the convolutional filters in order to enhance the conditioning of the gradients and therefore, help accelerate the convergence of the stochastic gradient descent optimization procedure. Formally, weight normalization reparameterizes the weight vector w as w = g. ||v|| v, where v is a new vector with Euclidean norm ||v|| and g is a scalar; now, with ||w|| = g (i.e. independent of the parameters v), the norm of the weight vector i.e. g is decoupled from the direction of the weight vector i.e. v/ ||v||. Consequently, stochastic gradient descent is performed directly in the new parameters v and g for better optimizability of the weights. We extracted the output from the n-th TCN block at the last time step, concatenated it with frames per stride and then parsed via a fully connected network to acquire the prediction output. Figure 3.5 visually details the TCN architecture consisting of 10 (hyperparameter) TCN residual blocks on the right, with the corresponding structure of a single TCN block in the middle and dilated causal convolution framework on the left.

3.2.4.2 Deeplearningclassifiers:Recurrentarchitectures

We applied 3 recurrent DL models for classification of gait strides. Similar to 3.2.4.1, Z-score normalized temporal features were provided straightaway as input for these classifiers.

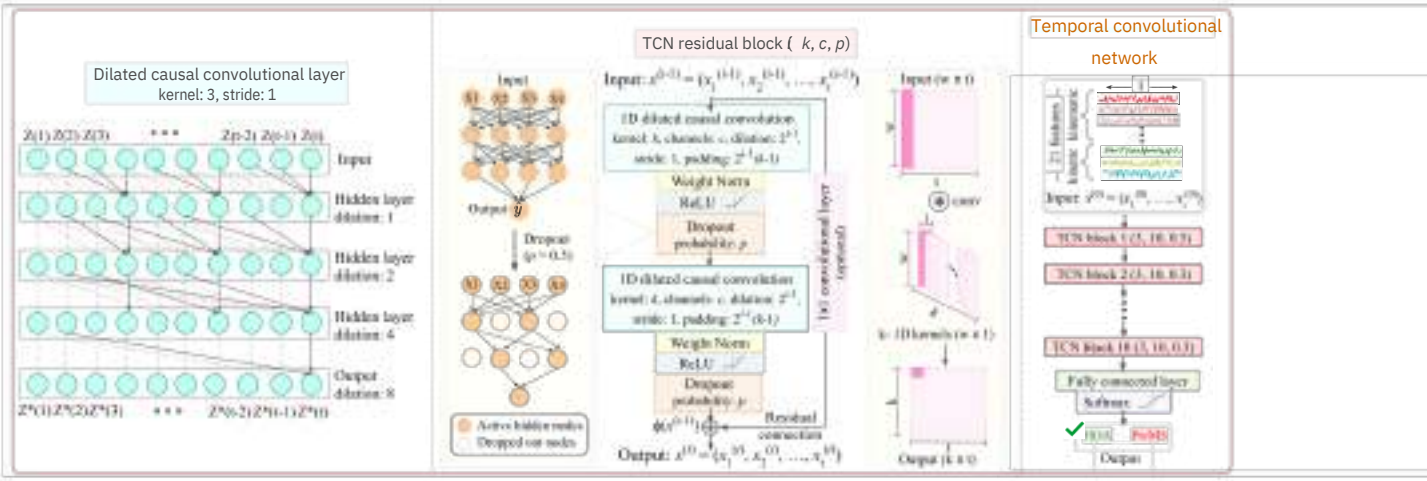

Figure 3.5: TCN architecture. Left: A dilated causal convolution with dilation factors d = 1, 2, 4, 8 and convolution kernel size of 3, convolution stride of 1, where Z(1), . . . , Z(t) being the input and Z ∗ (1), . . . , Z ∗ (t) being the output; Middle: A single TCN residual block, with x(i–1) being the input and x(i) as the output of the i-th TCN block; Right: A TCN of 10 blocks, connected with a fully connected end layer with softmax activation function, to generate the classification probabilities.

• Vanilla recurrent neural network (RNN): RNNs intrinsically integrate the sequential order of strides within a series as internal memory in their backbone architecture; this recurrence mechanism is not present in generic convolutional models. For recurrent layers, the output from the (t – 1)-th time step is fed back into the network along with the input at t-th step to determine step t's outcome. The corresponding gradients are computed using back-propagation through time. The output features from the last RNN layer at the last time step are provided to multiple fully connected layers to output the class prediction probabilities. Figure 3.6 schematically details a single RNN cell at the top left with input (xt), hidden (ht) and output (yt) states and a sample RNN architecture at the bottom left.

• Long short-term memory (LSTM): Although powerful temporal models, vanilla RNNs suffer from the vanishing gradient problem in longer sequences. That is, as we propagate forward in the network, small weight values for the hidden layers are multiplied together several times declining the gradients rapidly. Thus the weights for the initial layers are harder to train which in turn creates a domino effect for all further weights as well, making RNNs notably harder to train. LSTM [144], an extension of RNNs, mitigates these challenges via a memory cell with different gates to regulate the information flow into and out of the cell. Thus they are capable of handling long-short term dependencies in our gait stride inputs. Formally, an LSTM unit uses a cell state and 3 regulated gates, namely, input, forget and output gates, to add or remove information to the cell state.

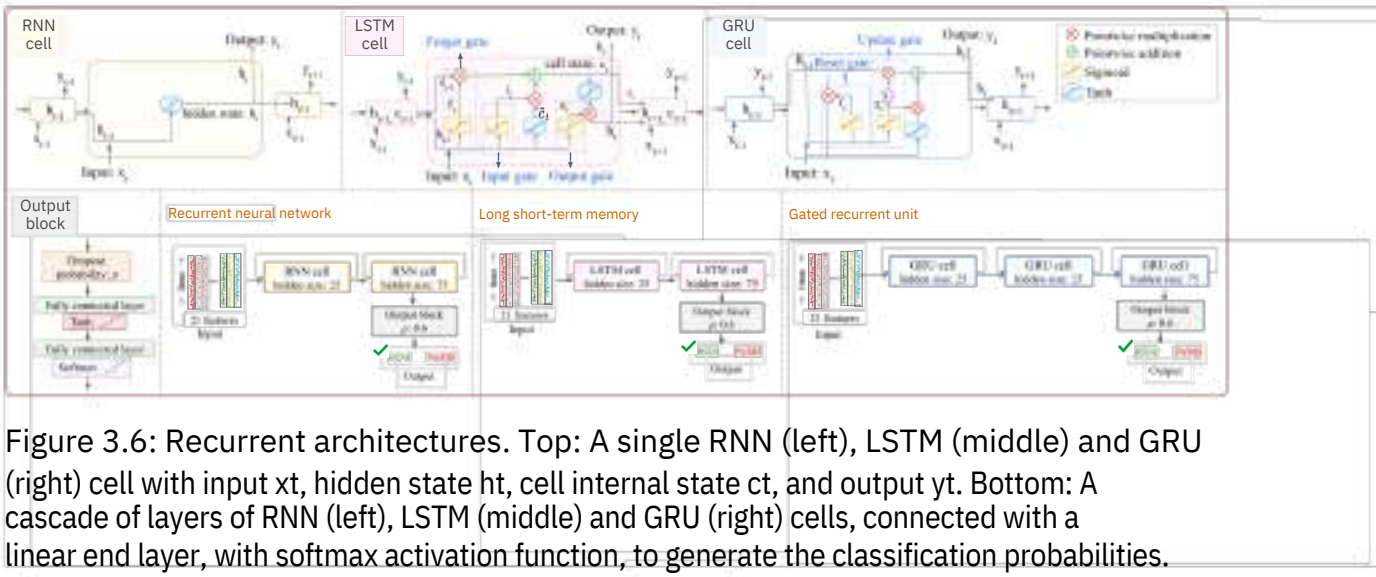

Figure 3.6: Recurrent architectures. Top: A single RNN (left), LSTM (middle) and GRU (right) cell with input xt, hidden state ht, cell internal state ct, and output yt. Bottom: A cascade of layers of RNN (left), LSTM (middle) and GRU (right) cells, connected with a linear end layer, with softmax activation function, to generate the classification probabilities.

Each gate consists of a sigmoid layer σ, with values between 0 and 1 describing the fraction of constituents to allow in via the gate, and a point-wise multiplication operation.

Forget gate ft = σ(Wf.[ht −1, xt] + bf) determines the information to discard from the cell state ct by looking at the current input xt and prior hidden state ht−1. Input gate it = Wi.[ht −1, xt] + bi) decides the values to update and the corresponding update to cell state $\tilde{c}_t$ is given by $\tilde{c}_t$ = tanh(Wc.[ht −1, xt] + bc). The new cell state is thereby computed as ct = ft ∗ ct−1+it ∗ $\tilde{c}_t$. Eventually, the output gate ot = σ(Wo.[ht−1,xt]+bo decides the portions of cell state to output with the output hidden state computed as ht = ot ∗ tanh(ct).

Our LSTM model had the same architecture as the RNN model described in 4.2.3.1, but with RNN layers replaced with LSTM layers (Figure 3.6). We experimented with both uni- and bi-directional LSTM layers. The extracted features from the n-th LSTM layer at the last time step are concatenated with the manually extracted feature and followed with a fully connected network to yield the output probabilities.

- Gated recurrent unit (GRU): Similar to LSTMs, GRUs [145] also use a gating mechanism to address the vanishing gradient issue. However, it eliminated the cell state and has only 2 gates, namely, reset and update gates; therefore, they have fewer parameters and are a bit faster to train than LSTMs. Update gate zt = σ(Wz.[ht −1, xt]) selects the information to add and discard in the hidden state, and reset gate rt = σ(Wr.[ht −1, xt]) determines on how much prior information to forget based on the current input xt and past hidden state ht−1. Then, the updated hidden state ht is computed as follows:

$$\tilde{h}_t = \tanh(W.[r_t * h_{t-1}, x_t]), \quad h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Again, our GRU model had the same architecture as RNN and LSTM models, but with GRU layers (Figure 3.6).

3.2.4.3 Modeltrainingandevaluation Our ternary (HOA, PwMS or PwPD) classification was studied across four different designs,
namely, task-, subject- (W and WT) and task-subject generalization (Figure 3.1). All classifiers for task generalization were trained on 1008 (HOA: 334, PwMS: 341, PwPD: 333) gait strides in W and tested to categorize 1016 (HOA: 351, PwMS: 332, PwPD: 333; corresponding imbalance ratio being 1.0: 0.95: 0.95) strides in WT across 25 common subjects that undertook both W and WT trials. Since our data set was limited to 1380 (HOA: 658, PwMS: 389, PwPD: 333; corresponding imbalance ratio being 1.0: 0.59: 0.51) strides across 32 subjects in W trials and 1050 (HOA: 351, PwMS: 332, PwPD: 367; corresponding imbalance ratio being 1.0: 0.95: 1.05) strides across 26 subjects in WT trials, we used a 5-fold cross-validation mechanism in all classifiers for both subject generalization frameworks. Task-subject generalization also utilized 5-fold cross-validation where training splits consisted of samples from 1380 strides in W and, correspondingly, we validated on separate subjects (than in training) with samples from 1050 strides in WT. In order to prevent information leakage, we ensured that no same subject had strides split between training and validation folds. Further, given the imbalance ratio in W strides, we applied stratification in all our cross-validation setups to preserve the class distribution of the whole dataset in each generated fold.

The computations for this work were implemented on a 12GB NVIDIA Tesla P100 GPU using PyTorch v1.7.0 DL platform in Python 3.6. In all classifiers, we set a fixed random seed for reproducible results. In all DL algorithms, the last layer outputs zi for class i were conv $\sum$ ertedtonormalizedpredictionprobabilitiesiusingsoftmaxactivationfunctionasi p p= ezi/n j1e.,where $\sum^{n}_{=}$ lce=− $\sum^{n}$
algorithmcomputedthegradientsofthelossrelativetotheweightparametersforallthelayers. These gradients were used to iteratively update the weights via stochastic gradient descent optimization algorithm in order to minimize the loss function. We experimented with various optimizers including, stochastic gradient descent with and without momentum, adaptive moment estimation (Adam), Adam with decoupled weight decay (AdamW) and root mean square propagation (RMSprop), all with varying learning rate schedules and weight decay regularization. We processed our data in batches of 128 strides each and randomly shuffle training samples at every epoch to reduce bias. We used early stopping to decide optimal number of training epochs, i.e., our training stops if the validation set accuracy did not improve after patience (hyperparameter)

epochs. In addition to weight decay and early stopping, using dropout between network layers also helped prevent over-fitting in our models. Several weight initialization schemes, such as the Xavier [146] and Kaiming [147] schemes, were tested to assist with vanishing or exploding layer activation outputs. To manage possible disparity in scales at appending the frames per stride feature with the processed model features, we tried layer normalization [148] to normalize each feature to zero mean and unit variance. Further, to handle class imbalance in our W trials, we implemented weight balancing by increasing increased the weights for minority class samples while computing the loss function. Exploratory hyperparameter optimization was performed to determine optimal framework for each learning classifier. See Section B.2.1 in the Appendix for more details on hyperparameter exploration.

In order to evaluate the prediction efficiency for the task generalization classifiers, we used the test set classification metrics, namely, precision, recall, accuracy, F1 score and area under receiver operating characteristic curve (AUC), whereas for the subject- and task-subject generalization frameworks, we computed mean and standard deviation in cross-validation metrics. Specifically for our multi-class evaluation setup, we used macro-averaging scores that averages the metric independently for each class, thus, treating all classes equally. All models were evaluated at stride- and subject-level categorizations, where majority voting was used to classify subjects as HOA, PwMS or PwPD. Thus a correctly classified subject's video had majority of strides accurately detected as of the appropriate cohort. Precisely, we annotate the stride and subject-level performance metrics with str (i.e. Pstr, Rstr, Astr, F1str, AUCstr) and sub (i.e. Psub, Rsub, Asub, F1sub, AUCsub) in the subscript, respectively. Further, for all DL models, we monitor learning curves for convergence of training accuracy and cross entropy loss metrics across epochs.

3.3 Results

3.3.1 Prediction Models

Overall, 16 classifiers were compared to categorize strides and subjects between HOA, MS and PD cohorts for task (3.3.1.1), subject (3.3.1.2) and task-subject (3.3.1.3) generalization.

3.3.1.1 Taskgeneralization

Table 3.1 summarizes the stride- and subject-wise evaluation metrics for top-3 ML and DL task generalization classifiers on categorizing the test set strides of trial WT. The aggregated perfor-

Table 3.1: Task generalization: comparing stride- and subject-wise test set performance across top-3 ML and DL algorithms

		Stride-based evaluation metrics					Subject-based evaluation metrics				
	Algorithm	Accuracy	Precision	Recall	F1score	AUC	Accuracy	Precision	Recall	F1score	AUC
ML	LSVM	0.781	0.784	0.780	0.780	0.905	0.960	0.963	0.961	0.963	1.0
	XGBoost	0.831	0.834	0.830	0.830	0.944	0.920	0.926	0.920	0.926	1.0
	GBM	0.831	0.835	0.830	0.824	0.945	0.960	0.963	0.961	0.963	1.0
DL	ResNet	0.825	0.834	0.823	0.876	0.972	1.0	1.0	1.0	1.0	1.0
	MSResNet	**0.896**	**0.903**	**0.898**	**0.899**	**0.97**5	1.0	1.0	1.0	1.0	1.0
	GRU	0.876	0.873	0.876	0.862	0.961	1.0	1.0	1.0	1.0	1.0

Note: ML is machine learning, DL is deep learning, AUC is area under the receiver operating curve; the numbers in bold represent the highest model performance. See Section 3.2.4 for details on algorithms.

mance of all the subject's strides via majority voting improved upon the accuracy of individual stride-wise predictions, for instance from 83.1% to 92% on XGBoost. The classification performances were higher across all metrics with DL algorithms than with traditional ML models. The top-3 DL models, viz. ResNet, MSResNet and GRU, all had perfect accuracy for classifying individuals with a given gait disorder (Asub) and the corresponding stride-level accuracy (Astr) of 87.6%, 89.9% and 86.2% (resp.). In contrast, the top-3 ML models i.e. LSVM, XGBoost and GBM, all resulted in an Astr of less than 85%. Analogously, the highest stride-level F1 (F1str) was 0.90 using MSResNet followed by 0.88 and 0.86 by ResNet and GRU (resp.), whereas F1str was lower than 0.85 applying any traditional approach. In Table 3.1, MSResNet had the highest accuracy, F1 and AUC of 89.9%, 0.90 and 0.98 (resp.) at stride-level, followed by ResNet and GRU with a matching perfect subject-level classification. The top task generalization algorithm was MSResNet trained for 40 epochs (as determined by the early stopping paradigm with pa- tience 10) with a batch size of 128, AdamW optimizer along with a learning rate of 0.002 and a weight decay of 0.01; with nearly 2.1 million model parameters, this model took 45 minutes (min) to train and 10 seconds to evaluate on a GPU. MSResNet uses both residual and multi- scaled learning framework to learn robust features from joint positions, even in motion. AdamW extends Adam optimizer to decouple weight decay from optimization steps, thus, the L2 regularization term is only proportional to the weight itself. This typically helps models to train faster and generalize better.

3.3.1.2 Subjectgeneralization

Table 3.2 summarizes the mean and standard deviation of 5-fold cross-validation performance metrics for the top-3 ML and DL subject generalization classifiers across W and WT trials independently. Similar to Table 3.1, the subject-wise diagnostic performance is higher than the

Table 3.2: Subject generalization: comparing stride- and subject-wise mean cross-validation performance across top-3 ML and DL algorithms

			Stride-based evaluation metrics					Subject-based evaluation metrics				
	Algorithm		Accuracy	Precision	Recall	F1score	AUROC	Accuracy	Precision	Recall	F1score	AUROC
W	LR	MLDT	0.576 ± 0.07	0.565 ± 0.06	0.558 ± 0.06	0.542 ± 0.05	0.732 ± 0.05	0690 ± 017.	0711 ± 018.	070021.	0671 ± 019.	0806 ± 013.
			0.557 ± 0.07	0.550 ± 0.06	0.532 ± 0.07	0.517 ± 0.06	0.691 ± 0.05	0633 ± 013.	0611 ± 015.	0611 ± 022.	0574 ± 017.	0844 ± 010.
		MLP	0.541 ± 0.05	0.530 ± 0.04	0.528 ± 0.05	0.514 ± 0.05	0.678 ± 0.04	0595 ± 011.	0589 ± 011.	0589 ± 010.	0558 ± 012.	0783 ± 008.
	DL	CNN	0.547 ± 0.07	0.526 ± 0.05	0.526 ± 0.07	0.506 ± 0.07	0.700 ± 0.07	0752 ± 011.	0722 ± 012.	0638 ± 019.	0647 ± 015.	0810 ± 012.
		ResNet	0.523 ± 0.05	0.503 ± 0.05	0.504 ± 0.05	0.492 ± 0.05	0.680 ± 0.05	0781 ± 021	0789 ± 022	0767 ± 025	0758 ± 024	0869 ± 011
	MS	ResNet	0.544 ± 0.08	0.501 ± 0.07	0.498 ± 0.08	0.492 ± 0.08	0.708 ± 0.06	0686 ± 014.	0656 ± 015.	0644 ± 020.	0622 ± 016.	0724 ± 012
WT	DT		0.516 ± 0.07	0.525 ± 0.08	0.521 ± 0.06	0.501 ± 0.08	0.643 ± 0.05	0650 ± 013.	0633 ± 012.	060015.	0580 ± 013.	0917 ± 006
	MLRF		0.514 ± 0.12	0.532 ± 0.11	0.508 ± 0.13	0.489 ± 0.13	0.707 ± 0.13	0567 ± 017.	0567 ± 017.	0538 ± 025.	0514 ± 022.	080020.
		MLP	0.546 ± 0.17	0.557 ± 0.15	0.547 ± 0.16	0.523 ± 0.18	0.734 ± 0.16	0683 ± 027.	070024.	0667 ± 034.	0640 ± 031.	0842 ± 023.
	DL	CNN	0.486 ± 0.12	0.479 ± 0.12	0.488 ± 0.11	0.470 ± 0.13	0.663 ± 0.12	0750 ± 017	0767 ± 013	0711 ± 021	0707 ± 018	0771 ± 017.
		MSResNet	0.513 ± 0.07	0.523 ± 0.07	0.503 ± 0.06	0.482 ± 0.06	0.709 ± 0.05	0720 ± 008	070 ± 012	0644 ± 018	0631 ± 014	0825 ± 009.
		GRU	0.522 ± 0.10	0.503 ± 0.08	0.519 ± 0.08	0.489 ± 0.09	0.687 ± 0.08	0633 ± 020	0633 ± 016	0589 ± 027	0571 ± 022	0725 ± 016.

Note: W is walking trial, WT is walking-while-talking trial; the numbers in bold represent the highest model performance in W and WT.

stride-wise measures. The top DL model, viz, ResNet for W trials and CNN for WT trials, outperformed all classical ML classifiers across all subject evaluation metrics in Table 3.2, except AUC for WT trials. Interestingly, none of the recurrent models made it to top-3 DL algorithms for subject generalization in W. The highest-performing subject generalization algorithm for W trials was ResNet with mean accuracy, F1 and AUC of 78.1%, 0.76 (class-wise F1: (HOA: 0.87, PwMS: 0.8, PwPD: 0.7)) and 0.87 (resp.), at subject-level. However, the top-3 ML models, namely, LR, DT and MLP, all ended up with a mean Asub, F1sub and AUCsub of less than 70%, 0.70 and 0.85 (resp.). Our highest-performing ResNet architecture employed positional encoding layer followed by an initial ConvBlock and 3 basic residual blocks, first with 64 filters and subsequent two with 128 filters, each with 2 ConvLayers with stride 1 and kernel sizes 8 and 5 (resp.). It was trained for 13 epochs with a batch size of 128 and AdamW optimizer (learning rate: 0.22 ×10−3, weight decay: 0.01); with nearly 360K model parameters, training took around 15 min on GPU. Further, to tackle imbalance, we weighed our loss function by 0.18, 0.36 and 0.45 for strides in HOA, PwMS and PwPD (resp.). Correspondingly, the highest-performing algorithm for subject generalization in WT was CNN with mean Asub, F1sub and AUCsub of 75%, 0.71 (class-wise F1: (HOA: 0.8, PwMS: 0.9, PwPD: 0.6)) and 0.77 (resp.). The top-3 ML mod-

els, namely, DT, RF and MLP, all had mean Asub and F1sub less than 70%, 0.65 (resp.), however surprisingly, DT had the highest mean AUCsub of 0.92. Our tuned CNN architecture had 2 ConvBlocks, first one having a ConvLayer with 64 filters of length 3 and stride 1 followed by batch normalization, ReLU and dropout layer with probability p = 0.4 and next one with a ConvLayer with 128 filters of length 2 and stride 1 followed by just ReLU activation layer. This CNN was trained for 25 epochs (35 min, 86K parameters) with 128 samples per batch and Adam optimizer with learning rate 0.001; no weight balancing was done in this case.

3.3.1.3 Task-subjectgeneralization

Table 3.3 summarizes the mean and standard deviation for stride- and subject-wise evaluation metrics of 5-fold cross-validation across top-3 ML and DL task-subject generalization classifiers. The top-3 DL models, i.e., CNN, ResNet and MSResNet, attained mean Asub of 79.3%,

Table 3.3: Task-subject generalization: comparing stride- and subject-wise mean cross-validation performance across top-3 ML and DL algorithms

		Stride-based evaluation metrics					Subject-based evaluation metrics				
	Algorithm	Accuracy	Precision	Recall	F1score	AUROC	Accuracy	Precision	Recall	F1score	AUROC
ML	LR	0.458 ± 0.12	0.485 ± 0.09	0.459 ± 0.12	0.450 ± 0.12	0.663 ± 0.10	0500 ± 33.	0500 ± 33.	0450 ± 33.	0467 ± 33.	0.706 ± 0.18
ML	AdaBoost	0.447 ± 0.16	0.468 ± 0.12	0.460 ± 0.16	0.441 ± 0.16	0.667 ± 0.12	0577 ± 29.	0600 ± 31.	0578 ± 32.	0564 ± 30.	0.732 ± 0.19
ML	MLP	0.505 ± 0.09	0.506 ± 0.10	0.505 ± 0.09	0.486 ± 0.10	0.679 ± 0.08	0507 ± 16.	0522 ± 17.	0444 ± 23.	0438 ± 20.	0.710 ± 0.09
DL	CNN	0.557 ± 0.08	0.567 ± 0.06	0.557 ± 0.08	0.545 ± 0.08	0.718 ± 0.07	0793 ± 024	0811 ± 025	0789 ± 29.	0782 ± 027	0.933 ± 0.11
DL	ResNet	0.538 ± 0.04	0.589 ± 0.04	0.547 ± 0.05	0.523 ± 0.05	0.747 ± 0.05	0707 ± 026	0756 ± 025	0694 ± 30.	0689 ± 028	0936 ± 007
DL	MSResNet	0.561 ± 0.09	0.612 ± 0.06	0.566 ± 0.07	0.552 ± 0.08	0.748 ± 0.06	0753 ± 021	0789 ± 019	0822 ± 0.16	0760 ± 020	0.922 ± 0.09

Note: The numbers in bold represent the highest model performance.

70.7% and 75.3% (resp.), and mean F1sub of 0.78, 0.69 and 0.76 (resp.). The top-3 ML models, namely, LR, AdaBoost and MLP, all had a mean Asub and F1sub of less than 60% and 0.60 (resp.). A 1D CNN had the highest overall subject-wise performance for task-subject generalization with the mean Asub, F1sub and AUCsub of 79.3%, 0.78 (class-wise F1: (HOA: 0.9, PwMS: 0.9, PwPD: 0.63)) and 0.93 (resp.). This optimal CNN had positional encoding followed by 3 ConvBlocks, each having a ConvLayer with 64, 128 and 64 filters (resp.), of corresponding lengths 9, 5 and 3 and stride 1 each. Further, batch normalization and dropout with p = 0.4 were used in the first ConvBlock and max pooling with kernel size 2 was applied in the last ConvBlock to manage overfitting. It was trained for 20 epochs (10 min) with RMSprop optimizer (learning rate: 0.001) processing 128 samples per batch with loss function weighed by 0.1, 0.35 and 0.55 for samples belonging to HOA, PwMS and PwPD (resp.). The model had total 86K parameters. It is interesting to note that convolutional models were top-performers across all designs.

3.3.2 Post hoc Analysis

Next, we 1) perform an ablation study to quantify the value of features from different body areas (3.3.2.1), 2) analyze feature importance (3.3.2.2) and 3) assess our DL predictions relative to physical performance of subjects (3.3.2.3).

3.3.2.1 Ablationstudy

We compared the task-, subject- and task-subject generalization performances on features from several body subsets, i.e, 18 (= 2 (left, right) × 3 (2 toes, 1 heel) × 3 (x, y, z)) feet-extracted (F), 24 combined feet-ankle (F+A) and 30 feet-ankle-knee (F+A+K) coordinates, to that of using all 36 lower body features. Precisely, we studied the impact of eliminating body parts in turn as we descend from hips to feet. All ML and DL models were trained and tuned from scratch on these data streams for comparison. Table 3.4 illustrates the stride- and subject-wise accuracy and F1 for the best performing algorithm on each data subset in task generalization; likewise, Table 3.5 reports Asub and F1sub for the highest-performing model on each subset with subject- and task-subject generalization schemes. DL models, specifically, CNN, ResNet and MSRes-

Table 3.4: Ablation study in task generalization

Data stream	Top-performing algorithm	Stride-based		Subject-based	
		A_{str}	$F1_{str}$	A_{sub}	$F1_{sub}$
F	MSResNet	0795	0796	0960	0958
F+A	ResNet	0828	0828	1.0	1.0
F+A+K	MSResNet	0892	0891	1.0	1.0
All	MSResNet	0899	0899	1.0	1.0

Note: F is feet, A is ankle and K is knee features; the numbers in bold represent the highest model performance.

Net, surpassed conventional ML performance across all data streams and model designs. Not surprisingly, the same three convolutional models were highest performers in 3.3.1 as well. Task generalization revealed the top stride-wise performance when using all 36 features with MSRes-Net (Astr: 90%), closely followed by F+A+K also with MSResNet (Astr: 89%) and then, F+A with ResNet (Astr: 83%); although, all data subsets, except using only feet features, had comparable subject-wise metrics. For subject generalization in both W and WT trials, using all features resulted in the highest mean cross-validation accuracy (Asub in W: 78%, WT: 75%) followed

Table 3.5: Ablation study in subject- and task-subject generalization frameworks

	Subject generalization (W)			Subject generalization (WT)			Task-subjectgeneralization		
Data stream	Top-performing algorithm	Asub	F_{1sub}	Top-performing algorithm	Asub	F_{1sub}	Top-performing algorithm	Asub	F_{1sub}
F	ResNet CNN	0629 ± 012.	0508 ± 011.	CNN ResNet	0583 ± 011.	0523 ± 009.	CNN	0707 ± 020.	0620 ± 023.
F+A	CNN ResNet	0624 ± 008.	0483 ± 007.	CNN CNN	0650 ± 024.	0616 ± 027.	MSResNet	0713 ± 025.	0674 ± 029.
F+A+K		0662 ± 010.	0620 ± 011.		0683 ± 011.	0627 ± 018.	CNN CNN	0723 ± 015.	0673 ± 018.
All		0781 ± 021	0758 ± 024		0750 ± 017	0707 ± 018		0793 ± 024	0782 ± 027

Note: F is feet, A is ankle and K is knee features; the numbers in bold represent the highest model performance.

by F+A+K (W: 66%, WT: 68%) and F+A (W: 62%, WT: 65%). A similar trend was noted in task-subject generalization where employing all 36 coordinates achieved top Asub of 79% via CNN, succeeded by F+A+K at 72% also with CNN and F+A at 71% with MSResNet. In all model frameworks, adopting entire lower body coordinates outperformed any other considered combination. Further, we saw a consistent improvement in performance as we augment additional coordinates, i.e, F < F+A < F+A+K < All, where < denotes an increase in our defined metrics. This indicated the importance of adding feet features to our study as their use in solidarity represented a major chunk of the overall model performance, for instance, Asub = 71% using only feet in comparison to 79% with all features in task-subject generalization. In conclusion, these ablation results indeed support our decision to use all lower body features for prediction.

3.3.2.2 Analysisoffeatureimportance In an attempt to explain and thereby establish trust in our classifications from DL models, we
examined global (via permutation feature importance) and local (via Shapley additive explanations (SHAP)) feature importance for our top models. Local feature importance focused on understanding the contribution of factors that led to a specific prediction, while global feature importance took all predictions into account.

- Permutationfeatureimportance: Permutationfeatureimportancemeasuredthedecrease in performance of our leading and optimally tuned DL algorithms, i.e., MSResNet for task generalization, ResNet and CNN for subject generalization in W and WT, respectively, and again, CNN for task-subject generalization, as we shuffle (x, y, z) position values of an individual body part, such as right knee and left heel. This permutation cuts the association between actual feature values and the corresponding class labels. Thus a lower performance after shuffling signified the dependence of our model on the associated fea-

ture for classification and consequently, a greater importance of the respective body part. We repeated our permutation process for the test set 10 times and averaged metrics over these repetitions for a robust result. Note that this does not involve any retraining of our top models. Figure 3.7 plots the AUCstr after permuting features relative to each body part for the optimal task generalization model i.e. MSResNet. Both knees followed by heels

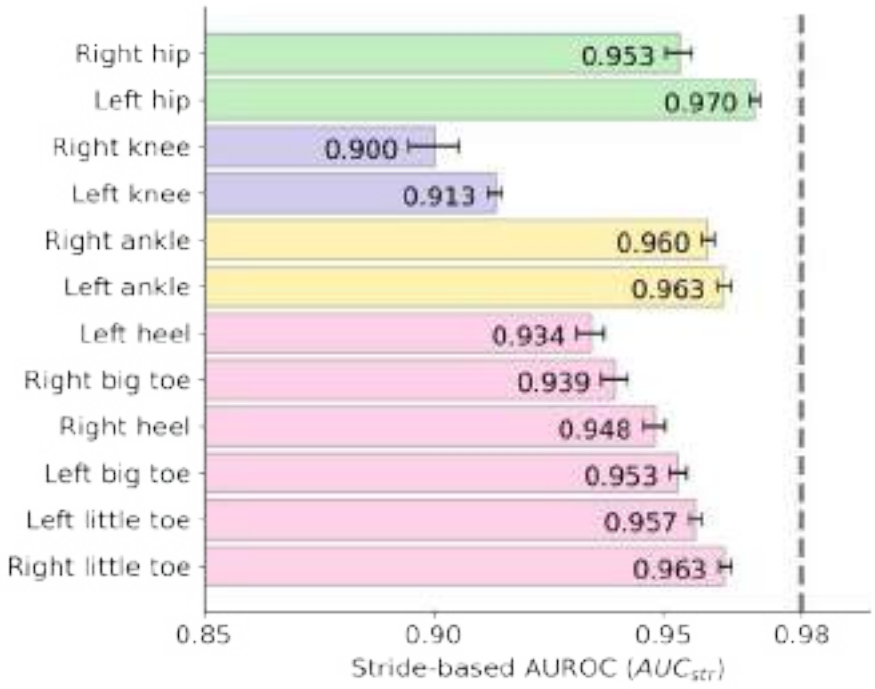

Figure 3.7: Permutation feature importance in task generalization. A low score (relative to best AUCstr of 0.98) after permutation signifies more importance. Hip, knee, ankle and feet keypoints are grouped in green, violet, yellow and pink (resp.), where features are sorted in decreasing order of importance within each group.

and big toes were the most informative features with the least AUCstr after permutation; however, left hip positions were the least predictive of labels. Overall, we observed that right-side features were more dominant than their left-side counterparts, for instance, features from the right knee were more important than left knee. Figure 3.8 plots the ratio of AUCsub after shuffling with respect to the original AUCsub of the subject- and task-subject generalization models. On average, right knee and hip followed by both little toes were the most relevant features, whereas, right ankle was the least important. It was interesting to observe that there were a few features, namely, right big toe, ankle and left knee, that had little effect on model performance for subject generalization in WT. In task-subject generalization (green triangles), all features seemed to be highly important as permuting any body part resulted in a loss of significant chunk in accuracy. This might indeed oc-

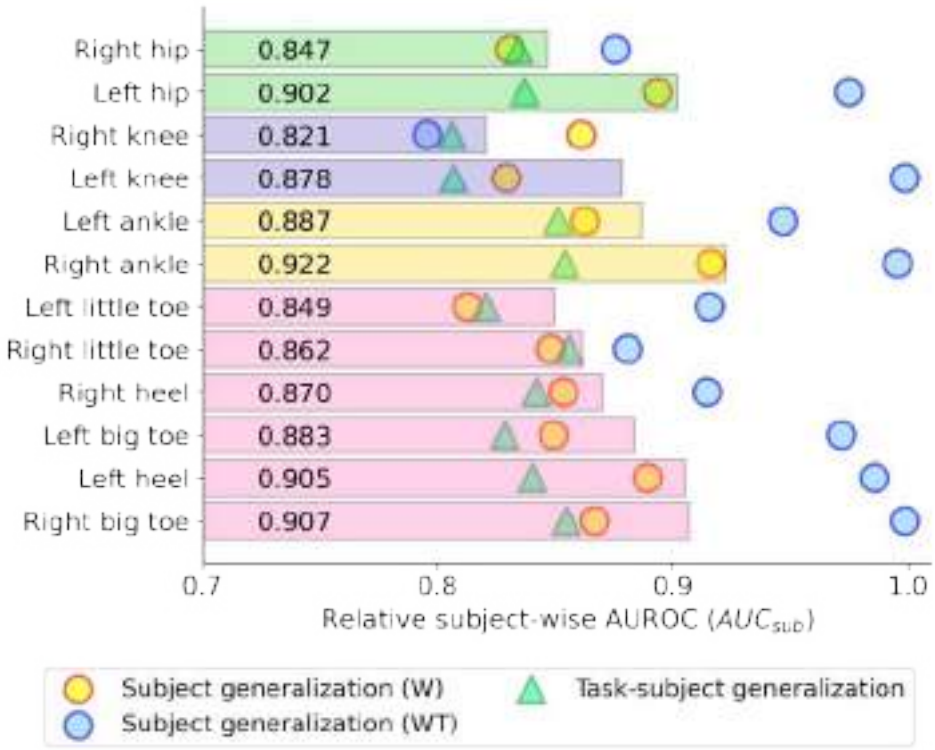

Figure 3.8: Permutation feature importance in subject- and task-subject generalization. The ratio of model's AUCsub after permuting a feature relative to its original AUCsub (0.87, 0.77 and 0.93 for subject generalization in W, WT and task-subject generalization (resp.)). Yellow/blue circles and green triangles denote the ratio in subject generalization W/WT and task-subject generalization resp; bars depict the average ratio across the three designs. A lower ratio indicates higher importance. Hip, knee, ankle and feet keypoints are grouped, where features are sorted in decreasing order of importance within each group.

cur due to it being a highly complex classification paradigm and therefore, all features together were essential to diagnose the heterogeneity present in new subjects in an unseen trial. Altogether, knee coordinates followed by several feet features seemed to be the most important for our analysis.

• Shapley additive explanations: SHAP [149] is based on a classic notion in game theory for optimal credit allocation, namely, Shapley values, where our classifier is considered analogous to a multi-player cooperative game with features as different players interacting together to produce the classification label as an outcome. Thus it is a local model-agnostic approach to assess feature importance by computing fair contribution of each player in our game. This kind of local explainability helps to understand individual stride characteristics that led to an accurate or erroneous prediction, which is indeed vital to facilitate targeted interventions in a medical setting. Figure 3.9 applies a SHAP decision plot to depict the

highest performing task-subject generalization model's (CNN) output trajectory for a sin- gle test-set stride that was correctly anticipated as belonging to a PwMS. The x-axis of

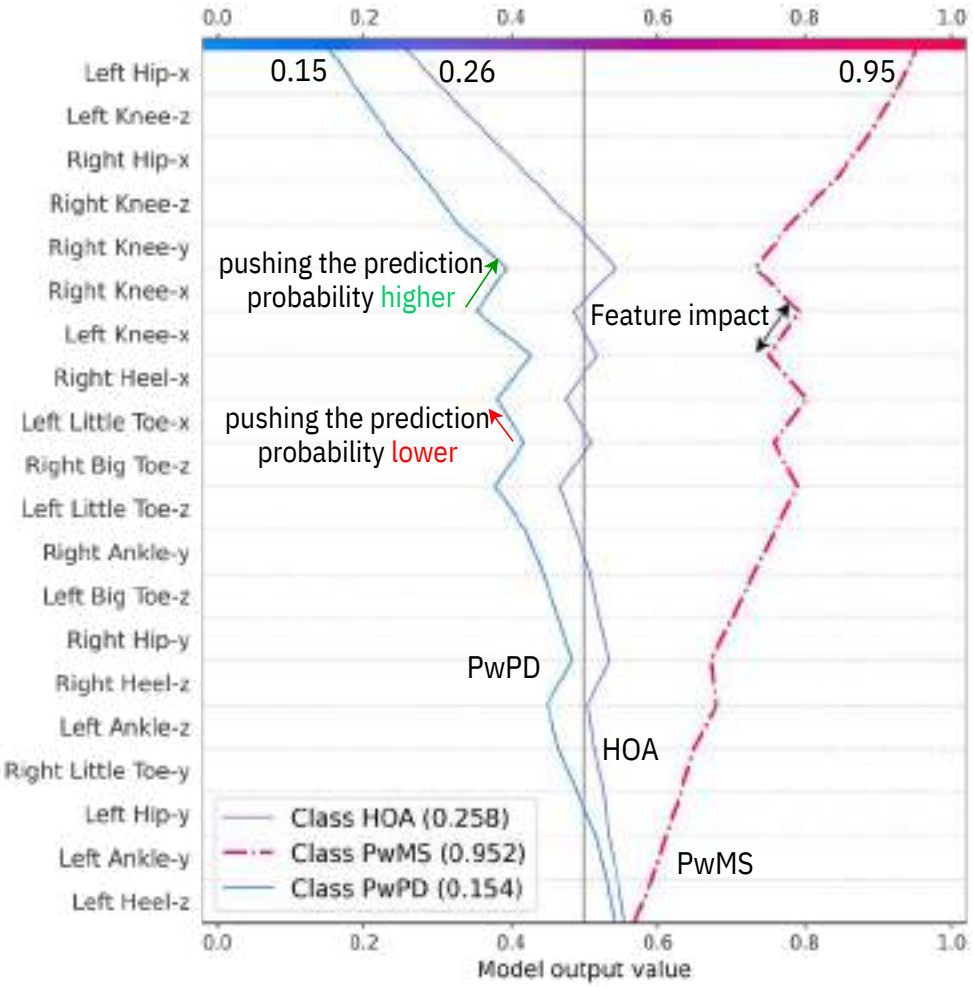

Figure 3.9: SHAP for the top-performing task-subject generalization model. Multi-output decision plot for a randomly selected stride, that was correctly classified to belong to a PwMS.

this plot is the model's output value illustrating the probability of stride getting classified (vs. not) as HOA, PwMS or PwPD; the y-axis lists the model's top-20 features in the descending order of importance. Note that this importance is calculated only over the stride examined. SHAP essentially evaluated the affect on model performance in presence vs. absence of each feature. Moving from bottom to top, SHAP values for each feature drove the model's output from the base value (average model output over the training samples) to the overall prediction output; features pushing the model output higher increased the class prediction probability and otherwise. Observe that the outlined stride was correctly classified as from MS cohort with the highest predicted probability of 0.95, via an aggregated impact from nearly all features. This matched our remark in permutation fea-

ture importance where all body parts together were critical for task-subject generalization. Moreover, knee coordinates seemed to dominate top features in Figure 3.9, similar to what we had observed in permutation feature importance. These model explanations, fit even to complex DL models, are crucial to ease the standard trade-off between accuracy and interpretability.

3.3.2.3 Associationwithlowerextremityfunction We attempt to inspect a potential association between our top DL model predictions and the
corresponding lower extremity physical function of subjects. We used the short physical performance battery (SPPB) assessment [150] as a common measure to evaluate the lower extremity functioning in all our older adults. SPPB integrates the performance of subjects in gait speed, chair stand and balance tests to create a summary score between 0 (worst) and 12 (best); lower scores indicate severe mobility limitations and higher scores indicate better performance. We had subjects with minimal to moderate frailty, i.e. SPPB: 9.85 ± 2.35 [6 – 12], in our data.

Figure 3.10 visualizes the predictions made by the highest performing subject generalization in W model (ResNet) with respect to the frailty level in corresponding subjects, as measured by SPPB. The markers, i.e. circles, squares and triangles, represent actual HOA, PwMS and PwPD,

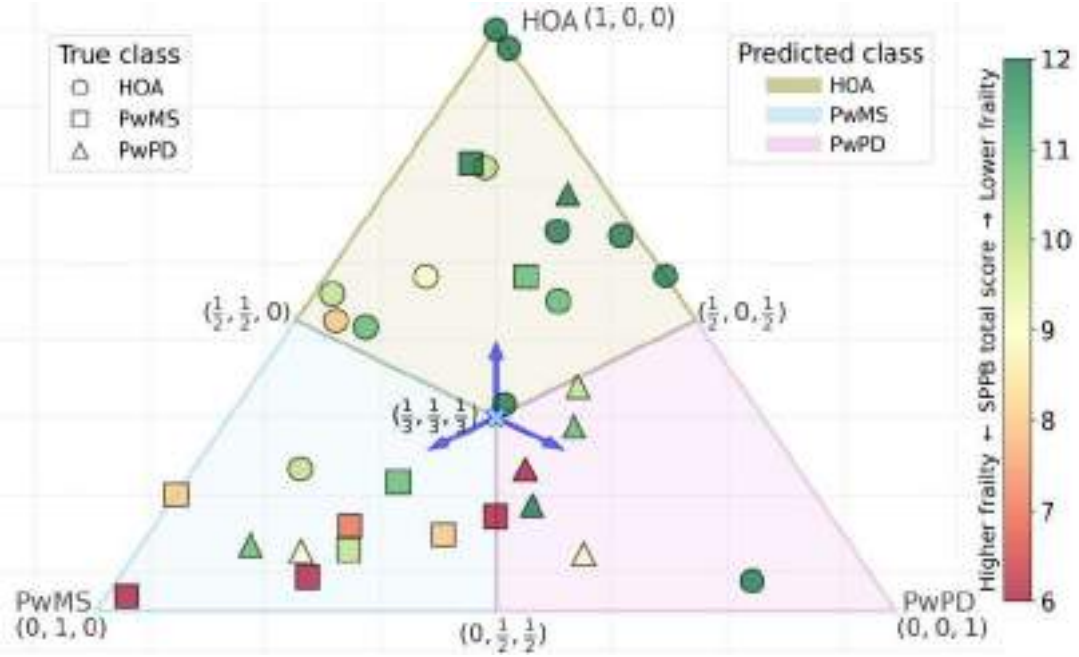

Figure 3.10: Visualizing the predictions of subject generalization (W) model with respect to the corresponding lower extremity strength of subjects.

respectively, where marker face-color shades denote the subject's SPPB. Moreover, marker po-

sitions are the barycentric coordinates representing the 3D predicted probability vector on an equilateral triangle. Consequently, our triangle is trisected in 3 equal parts, where background color depicts the predicted class by model, i.e, markers on green, blue and purple segments are predicted as HOA, PwMS and PwPD, respectively; and centroid represents an equal probability for each class. Therefore, a perfect classification would correspond to all circles in the north, squares in the south-west and triangles in the south-east vertex of the triangle. Not surprisingly, we observed that HOAs with a lower SPPB (higher frailty) had greater number of strides mis- classified as belonging to MS cohort and likewise, PwMS with a higher SPPB (lower frailty) had majority of strides incorrectly predicted as from the healthy cohort. We also see from Figure 3.10 that it is hard to distinguish between MS and PD subjects, which was again expected given the existent heterogeneity in the disorders. In summary, SPPB seems somewhat correlated with our model predictions but is not solely enough to distinguish among classes.

3.4 Discussion

This study proposed a novel framework using multi-view visual data driven DL for MS and PD gait dysfunction prediction. Our system provides a convenient, low-cost, accurate, and rapid remote monitoring tool for neurological gait classification. Our architecture does not need any certified professional in charge, and being contact-less, it provides convenience and automatic- ity in the gait assessments of older adults in the wild. Our workflow is end-to-end open source,
available at A few other works
have explored vision data to categorize neurological gait [119, 127]. In contrast to our comprehensive comparison with 16 diverse models across four different designs, namely, task-, subject- (W and WT) and task-subject generalization, others [119, 127] have only examined CNN and ResNet, respectively, for a subject generalization in W trials. Our work thoroughly explored the interpretability of optimal models via post hoc analysis, which was missing in past studies. The presence of body markers in other research [127] might overfit to solely detected markers; and one study [119] performed manual feature engineering after the extraction of 2D positions, which is now automated in our work. In comparison with prior work examining the binary clas- sification of PwMS versus only controls using wearable-derivable measures [72], where 80% accuracy has been achieved when generalizing across new subjects, the ternary classification ap- proach explored here, provides a proof of concept of a gait classification framework capable of identifying different origins of neurological gait disorders at a similar level of performance.

We observed that convolutional models were highest-performing across all generalization frameworks, which should provide guidance for future work in neurological gait classification.

However, no one specific architecture was found to yield the best performance across different features, tasks, and frameworks. These findings suggested the importance of exploring differ- ent DL architectures in future work examining gait to extract as much information as possible from the input data. From a clinical perspective, stride-wise classification allowed for the use of a single stride, or brief duration walking trial, to serve as the basis for disease monitoring, which might be well suited for clinical settings with limited space and time. We used OpenPose- extracted position coordinates as input to our DL models instead of raw images as trained models with the latter might be sensitive to subject's footwear, clothing, and background, whereas Open- Pose is robust to most of these factors. Our ablation study results demonstrated the importance of feet features in neurological gait classification, particularly in our task-subject generalization framework, which generalizes to new participants in different walking tasks. Further, through an analysis of permutation feature importance, we found the importance of right knee features, which might be partly due to the right-side dominance of participants in this study or positioning of video camera on right side of treadmill. SHAP visualizations (Figure 3.9) provided a com- pact and efficient view of our model explanations to highlight relevant features for practitioners. These interpretable explanations not only helped to understand, but also trust the findings from our system.

The current study explored an automated gait screening model but the small sample size and gender differences between groups recruited for this study limits making generalized interpretations. While 3D joint coordinate trajectories used for classification were not compared with a lab-based motion capture system, prior work suggests an accuracy of 30 mm or less with re- moval of failed body segment recognition [151]. Since we relied on cross-validation to gauge the performance of subject- and task-subject generalization models, evaluating on a holdout data set would be essential to establish robustness. Exploration and analysis on the optimal number of continuous strides needed for best results is a crucial next step. Future research should examine inclusion of more types of pathological populations and the effect of number and position of digital cameras on the performance. Finally, evaluating the utilization of a 3D body mesh instead of sparse 3D coordinates might help improve the pose estimation block of our system. Future work might also involve exploring recent hybrid intelligence-driven and graph neural network-based approaches [152, 153].

3.5 Summary

The expression of neurological conditions over time and aging is heterogeneous, making the identification of sudden changes in PwMS and PwPD particularly difficult. We presented a

CHAPTER 4

USING MULTI-STRIDE DYNAMICS IN GAIT

In this chapter, we review the work in Deep Learning for Multiple Sclerosis Differentiation Using Multi-Stride Dynamics in Gait. This work is currently in review as [74]. Our entire code for this work is publicly accessible.

4.1 Introduction

Multiple sclerosis (MS) is an immune-mediated, neurodegenerative disease that affects approximately 1 million people in the United States and more than 2.5 million globally [13, 84], with a shift in peak prevalence to adults 55-64 years of age [18]. MS can be immensely heterogeneous; persons with MS (PwMS) may suffer from extremely mild to severe muscle immobility, speech and vision complications, and memory issues [15]. Gait and balance dysfunction are common symptoms in PwMS, with nearly 85% of PwMS describing gait disorders as a major complication [9] and roughly 50% of patients needing walking assistance within 15 years of MS onset [20]. Gait performance declines have been observed in PwMS, particularly as disability increases [86, 92, 93, 154]. Past studies have found reduced gait speed, shorter steps, extended stride time, wider base of support, reduced single support phase, and a prolonged double support phase in PwMS compared to controls [92,93]. However, most gait-based methods for identifying MS have relied upon traditional statistical techniques to examine differences in spatiotemporal features and correlations with disability. Compared to statistical testing that analyze features individually, machine learning (ML) models are capable of utilizing linear and nonlinear combinations of spatiotemporal and kinetic gait features to potentially improve MS gait identification. Given the increased access to objective gait data from wearable technologies or traditional gait labs, supervised ML methodologies have been increasingly used in human gait analysis across neurological populations, including MS [72,94,95]. In particular, ML methods like random forest and artificial neural networks have been used to identify gait changes in Parkinson's disease in [94, 95], whereas [72] focused on MS-related changes. With the increasing successes of deep

learning (DL) across domains, recent works [113, 155] compared several ML models with the long short-term memory (LSTM) DL approach to distinguish between low and high fall risk in neurological gait. The LSTM model outperformed all traditional ML methods (classification accuracy: 0.94 (LSTM) vs. 0.88 (top ML model) in [155] and 0.86 (LSTM) vs. 0.73 (top ML model) in [113]), showcasing the potential of DL in human gait analysis. See [156, 157] for a more detailed comparison of ML and DL approaches in gait analysis.

This work attempts to examine MS related changes in spatiotemporal and kinetic gait features across multiple strides; and evaluate the effectiveness of deep learning for MS differentiation using multi-stride dynamics in gait (DeepMS2G). Specifically, we propose a DL-based method- ology to classify multi-stride sequences of PwMS from healthy controls (HC), so as to generalize across different walking tasks and subjects. Building upon prior work examining MS classifi- cation using traditional ML frameworks on individual strides [72], we categorized PwMS using the following 2 classification designs (see Figure 4.1):

1) Task generalization demonstrating the generalization over different tasks. Specifically, we train binary (healthy vs. MS) supervised classifiers on walking (W) trials and test them on walking-while-talking (WT) trials, to examine how findings from data collected in a clinic or lab may generalize to more realistic gait tasks.

2) Subject generalization establishing the generality over newer subjects. Specifically, we train binary classifiers on some subjects and apply them to an independent set of withheld test subjects.

Concretely, our contributions are as follows:

- WepresentedaDLapproachtodifferentiateMSrelatedchangesfromcontrolsusingmulti-stride dynamics in spatiotemporal and kinetic gait features. Of particular novelty is our focus on MS.
- We utilized multi-stride dynamics from 21 extracted kinematic and kinetic gait features.
- Webenchmarkthecomparativeperformanceof16diverseMLandDLmodelsforMSdif-ferentiation across two classification frameworks, i.e. task and subject generalization and two feature scaling strategies, i.e., body size- and multiple regression-based normalization.
- We investigated the explainability of our top-performing algorithms via ablation study on gait features and feature importance. Moreover, we discussed the association between the lower extremity function of participants and our model predictions. This post hoc analysis of DL models was absent in previous analogous studies.

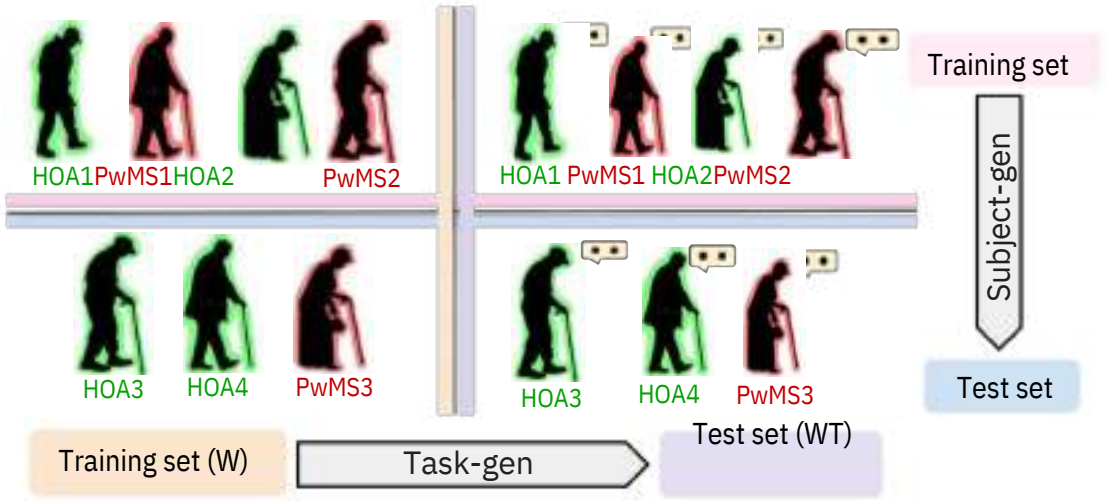

Figure 4.1: Task and subject generalization model designs. Healthy older adults (HOA) and PwMS are depicted in green and red highlights, respectively. The indices (1, 2, 3, . . .) are used to indicate dummy subject identifiers.

4.1.1 Related Work

Neurological gait disorders like MS are characterized by reduced mobility, abnormal gait mechanics, poor balance and muscle weakness, as well as cognitive and autonomic dysfunction [158, 159]. These symptoms typically lead to fatigue and physical inactivity and consequently increase the risk of development of secondary diseases. Several works on movement analysis have utilized wearable inertial measurement unit sensors [123], electromyography (EMG) [160], and motion capture systems [161] to predict neuromuscular changes in neurological gait. Past studies on gait-based methods for identifying MS have relied upon statistical significance tests such as t-test, and ANOVA (analysis of variance) to examine differences in average and vari- ability of spatiotemporal features, and correlations with neurological impairment assessed by Kurtzke's Expanded Disability Status Scale [86, 87, 90, 162]. Compared to the statistical tests that analyze features individually, ML and DL models are capable of determining multivariate discriminants taking into account multiple features. Further, these algorithms can also produce non-linear decision boundaries, potentially leading to superior accuracies. Recently, several studies have focused on traditional ML to classify gait patterns in PwMS [72, 97]. Addition- ally, authors in [113] used a long short-term memory model to distinguish between low and high fall risk in PwMS via accelerometer sensors. We utilized data driven DL for classification of multi-stride sequences of PwMS from HC utilizing domain knowledge-based spatiotemporal and kinetic gait features. Note that in comparison to [113], we studied a different classification task and used a separate cohort.

In this work, we utilized spatiotemporal and kinetic gait parameters as input features, which contain valuable domain knowledge with the potential to improve classification performance. Further, since the effect of gait normalization is seemingly unexplored in the existing MS literature, we compared the classification ability of all models with standard size-based and multiple regression normalization schemes, first explored in [95, 99], across both the studied task- and subject generalization model designs. Moreover, we explored the explainability of our top-performing algorithms; and discussed the association between the lower extremity function of participants and our model predictions. The proposed methodology is an advancement towards the development of a system aimed at integrating anthropometric and wearable-derivable data to differentiate MS-related changes in older adults, and provide clinicians or informal health care members to monitor sudden changes in gait function and disability in older PwMS. Our eventual goal is to automatically determine the onset of MS and provide an explanation of the automated diagnosis.

The remainder of the chapter is organized as follows. In Section 4.2.1, we introduce the data acquisition paradigm and study participants. In Sections 4.2.2 and 4.2.3, we discuss the data analysis methodology, including feature extraction, data normalization, and classification strategies. In Section 4.3, we present our model prediction results and post hoc analysis. Finally in Sections 4.4 and 4.5, we highlight some concluding remarks along with limitations and future directions for this study.

4.2 Methods

4.2.1 Design of Experiments: Subjects and Setup

In this section, we present the subject demographics (4.2.1.1) and the experimental protocol for our gait data acquisition (4.2.1.2); our study was approved under the University of Illinois institutional review board number 15674 on 4/3/2015.

4.2.1.1 Studyparticipants

The sample consisted of 40 subjects; 20 PwMS (age: 61.05 ± 6.87 years [49 – 75 years], male/female: 5/15) from the local community. Our inclusion criteria ensured all [48 participants were medically stable, i.e., a score of above 18 on the telephone interview for cognitive status [106], with no recent lower limb injury; further, subjects were right-side dominant

± 6.87 years
– 68 years], male/female: 5/15) and 20 age, weight, height and gender-matched HC (age: 61.2

and had corrected to normal vision. All included PwMS were relapse-free for the past month, had mild to moderate disability, i.e., 4.3 ± 1.62 [1.0 – 6.0] on the Kurtzke's Expanded Disability Status Scale (EDSS) [107], and had no other cognitive disorder that may additionally influence their body balance. Note that 2 HC and 3 PwMS were excluded for holding the handrails while walking on the treadmill and thus biasing their force readings. Also, we separately reserved a group of 30 additional HC (age: 67.6 ± 10.34 years [50 – 87 years], male/female: 9/21) to normalize our extracted gait features (see 4.2.2.2 for further details).

4.2.1.2 Experimentalparadigm We used a C-Mill, Motekforce Link instrumented treadmill for participants to walk at their self-paced speed during all experimental trials. The treadmill had a built-in force plate supporting kinetic data acquisition, specifically, allowing for vertical ground reaction forces to be collected. Each participant completed a 75 seconds trial at their self-selected pace under 2 configurations, i.e., single-task paradigm W and dual-task condition WT; participants recited alternate letters of the alphabet while walking for WT trials. All participants were instructed to equally prioritize their attention between gait motion and the cognitive exercise during WT trials. Past studies have demonstrated differentiation between the single- and dual-task designs, where WT (in comparison to W) illustrated more resemblance to real life daily gait in middle-aged to older adults [163]. Throughout each walk, the built-in treadmill software recorded 1) position coordinates and time stamps for each gait event, such as left and right heel strike, using a single force plate, 2) ground reaction forces, 3) treadmill speed, and 4) center of pressure position coordinates at 500 Hz.

4.2.2 Gait Feature Extraction and Designs

In this section, we introduce our data analytic approach to construct supervised DL classification models predicting PwMS using gait variability across multiple strides. In particular, we describe the gait feature extraction (4.2.2.1), data normalization (4.2.2.2), stride augmentation (4.2.2.3) and feature design (4.2.2.4) techniques. Our detailed gait data analysis pipeline is illustrated in Figure 4.2.

4.2.2.1 Gaitterminologyandfeatureextraction A stride or gait cycle has the following phases (in order), HSR: heel strike right, TOL: toe-off left, MidSSR: midstance right, HSL: heel strike left, TOR: toe-off right, MidSSL: midstance

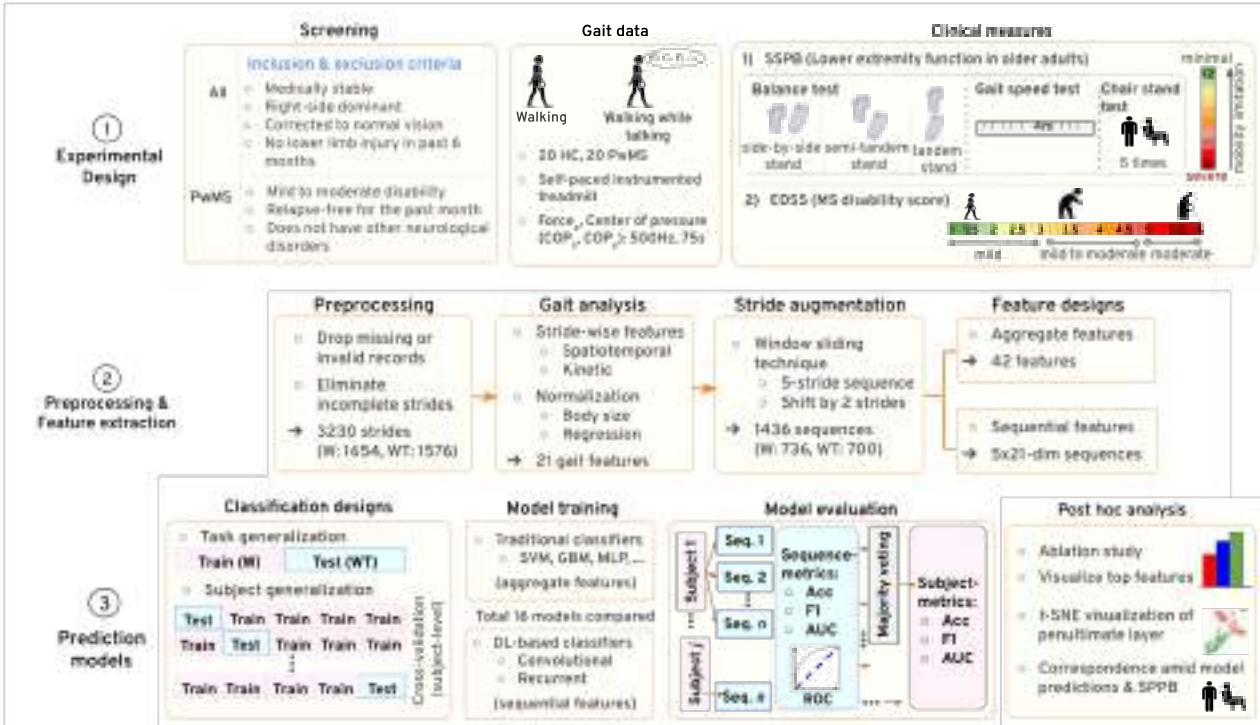

Figure 4.2: Workflowpipeline. Theproposed DeepMS2G (deep learning for MS differentiation using multi-stride dynamics in gait) framework.

left, and subsequent HSR beginning the next stride. We extracted 21 characteristic spatiotemporal and kinetic features across strides from the raw gait data to comprehend distinguishing patterns between HC and PwMS gait. See Figure 4.3. The extracted features can be organized

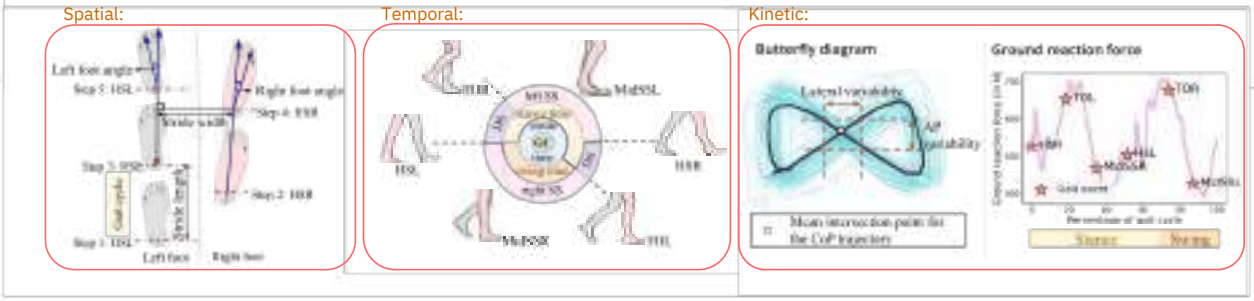

Figure 4.3: Gait features. Left: Spatial features, namely stride width, length and foot progression angles, Middle: Temporal features, namely stride, stance, swing, single support (SS) and double support (DS) times, Right: Kinetic features, namely butterfly diagram-based variability and forces. GC is gait cycle, AP is anterior-posterior, and CoP is center of pressure.

in following 4 categories:

• Spatial: 4 spatial gait features, i.e., stride width, stride length and the left and right foot progression angles [108], were extracted for each stride. Figure 4.3 (left) diagrammatically

summarizes the definition for these features on an overground view of the gait patterns. See [72] for detailed definitions of features.

• Temporal: 7 temporal features, i.e., swing time, stance time, stride time, supporting (right single, initial double and terminal double) times and cadence, were extracted for each stride. Figure 4.3 (middle) illustrates these features on a sagittal plane view of a stride starting at HSL.

• Spatiotemporal: 2 spatiotemporal markers, i.e., 1) stride speed defined as the ratio of stride length and stride time, and 2) walk ratio defined as the ratio of stride length to the count of strides covered in a minute, were extracted for each stride.

• Kinetic: 8 kinetic features, i.e., 6 forces, at each of the 6 gait events, and 2 butterfly diagram-based features, were extracted for each stride. A butterfly diagram [109] defines the recurrent center of pressure trajectory for several strides throughout a participant's walking trial. We extracted 2 characteristic gait features from the butterfly diagram, namely, 1) lateral shift in the intersection point of the center of pressure trajectory, and 2) lateral squared deviation from the average intersection point for a trial. See Figure 4.3 (right).

After eliminating nonconsecutive strides and those with missing or invalid gait events, we obtained 1654 (HC/PwMS: 905/749) and 1576 (HC/PwMS: 878/698) strides from W and WT trials (resp.), across 18 HC and 17 PwMS. Deriving these multiple samples per subject's walk significantly augmented as well as introduced variations to our data.

4.2.2.2 Datanormalization Similar to our prior work on PwMS [72] and past studies examining other neurological disorders
[95, 99], we compared the following two normalization approaches to reduce the intrinsic bias of our extracted gait features on the demographics of the subject and thus, improve the MS gait identification accuracy:

• Body size-based normalization (size-N): Gait features were normalized to dimensionless quantities via dividing by their corresponding matching dimension body size-based scaling factors [110]. Denoting the body weight (in kg), height (m), shoe size (m) and acceleration of gravity (9.81 m/s2) by w, h, Ssize and g, respectively, Table 4.1 summarizes the size-N gait features.

Table 4.1: Body size-based normalization

Raw gait parameter	Dimensionless quantity
Length: $L \in$ {stride length, stride width}	$\tilde{L} = \frac{L}{\sqrt{Thg}}$
Time: $T \in$ {stride time, stance time, swing time, supporting times}	$\tilde{T} = \frac{}{}$
Force (F_{ez}): 6 forces, one at each gait event e	$\tilde{F}_{ez} = \frac{F_e}{wg}$
Cadence: C	$\tilde{C} = \frac{}{}$
Stride speed: SS	$\tilde{S}S = \frac{SS}{\sqrt{gh}}$
Angle: $\theta \in \{\theta L, \theta R\}$	$\tilde{\theta} = \theta$
Walk ratio: W	$\tilde{W} = \frac{W}{\sqrt{h}g}$
Centerofpressure: $P \in \{\beta L, \alpha L\}$	$\tilde{P} = \frac{P}{Ssize}$

- Multipleregression-basednormalization(Regress-N):Weregressedthegaitfeaturesof normative walking data from 30 additional HC (age: 67.6 ± 10.34 years [50 – 87 years], male/female: 9/21) on multiple demographic characteristics and used these as baselines to normalize our extracted gait features. We derived the same 21 gait features from a total of 3923 valid strides obtained from our 30 additional HC. A regression model, which minimized the Tukey biweight loss of standard Gaussian residual errors, was fitted to each gait feature. In this regression, independent variables were the demographics (weight, height, gender and age); and dependent variables were subject-wise averaged gait feature values (as defined earlier in Section 4.2.2.1). Note that we only used these 30 additional HC that were not part of the main study (4.2.1.1). Gait features from both trials (W, and WT) of the main study subjects were then normalized to dimensionless quantities, where their predicted values were obtained via their corresponding fit and subject demographics. We analytically studied the efficiency of all models to classify MS gait with both the discussed normalization schemes, expecting the regress-N approach derived from multiple demographic properties to demonstrate an improved performance.

4.2.2.3 Strideaugmentation

Building upon past work on fall risk assessment [155], we followed the moving window method to assemble the extracted gait features from 5 time-consecutive strides, creating a 5 ×21-dimensional sequence (data sample) with 21 features (time series) over 5 temporally ordered

strides (time steps). Subsequently, we moved our window by 2 strides to devise the next 5-stride data sample. Thus we derived numerous multi-stride samples per subject, each capturing the gait variability and dynamics across 5 heterogeneous strides. This way, we substantially augmented and introduced variations to our original subject-level data in 4.2.1.1. This data augmentation approach might assist in the generality and training process of our complex DL models. Overall, we formed 736 (HC: 416, PwMS: 320) and 700 (HC: 399, PwMS: 301) 5-stride sequences from W and WT trials (resp.), across 35 subjects.

4.2.2.4 Featuredesigns

Next, we used our derived 5 × 21-dimensional samples to design 1D aggregated gait features vector and 2D sequential data, suited for our traditional ML- and DL-classifiers (resp.).

- Aggregated features: We used mean and standard deviation to aggregate our 5 ×21, i.e., 2D, sequences along the time dimension and construct a 1D feature vector of length 42, which is the expected input for any classical ML model like decision tree, etc. Thus we compiled a dataset of 1436 data samples across W and WT trials with 42 average- and deviation-based features per sample.
- Sequential features: We directly used the extracted 2D-sequences (5 × 21) as the input for all our convolutional as well as recurrent DL models. This 2D data encompasses both domain-knowledge along with temporal variations in subject's gait and further, did not risk losing information during aggregation. Overall, our input for DL models was 1436 samples, each consisting of a 21-channel sequence, with spatiotemporal and kinetic gait parameters, over 5 consecutive time steps, capturing possible dynamics in the gait data.

4.2.3 Classification and Evaluation

We examined binary classification to differentiate between 5-stride sequences of HC and PwMS across task- and subject generalization frameworks. Overall, we compared 16 models (see 4.2.3.1); in particular, 9 traditional ML algorithms, 4 convolutional and 3 recurrent DL architectures, across both classification frameworks, with corresponding model training and evaluation details in 4.2.3.2 and 4.2.3.3 (resp.). For task generalization, all models were trained on 736 5-stride sequences across 35 subjects in W trials and tested to categorize 700 sequences of the same subjects in WT trials. Given a limited dataset with 35 subjects, we used 5-fold cross-validation for subject generalization design. Further, all models were compared across both size-N and

regress-N normalized features. All features were Z-score normalized before inputting them to the model.

4.2.3.1 Classificationmodels Firstly, we examined 9 traditional ML algorithms: logistic regression (LR), support vector machine with linear (LSVM) and radial basis function (RBF SVM) kernels, decision tree (DT), random forest (RF), adaptive boosting (AdaBoost), eXtreme gradient boosting (XGBoost), gradient boosting machine (GBM) and multilayer perceptron (MLP) [164]. We used the aggregated gait features as input for these classical ML algorithms. Next, we compared the following 4 convolutional DL models (see 4.2.3.1 to 4.2.3.1) and 3 recurrent models (4.2.3.1 to 4.2.3.1); for these algorithms, a sequence of 5 consecutive strides was used directly as the model's input. These algorithms have been previously used for vision-based gait analysis in our past work [73].

- 1D Convolutional Neural Network (CNN): Our CNN architecture consisted of multiple convolutional blocks where each block was composed of a 1D convolutional layer succeeded by a batch normalization layer, non-linear activation function, dropout layer [136] and a pooling layer. The convolution function hierarchically extracted low-level features from the input data in the initial few convolutional layers to more complex high-level characteristics as subsequent layers are applied in the architecture. We experimented with several activation functions to introduce non-linearity into our convolutional layer output neurons, including a rectified linear unit (ReLU). We also explored dropout layers to randomly disable neurons and their corresponding connections to avoid over-fitting during the training process. Further, pooling was included to lower the number of model parameters, and ensure that convolutional layer-extracted features were invariant to minor translations in the input data. The output was then passed through multiple feed forward layers and finally, our final linear layer yielded a vector of length 2.

In contrast to recurrent DL models with an inherent sense of sequential processing for temporal data, CNNs (where the entire sequence is fed at once) may not necessarily handle strides within a multi-stride sequence relative to their positional order. Consequently, we used the sinusoidal positional encoding [137] to explicitly add this information to the input.

- Residual neural network (ResNet): ResNets learn residual functions relative to the layer inputs and thereby, assist in the training of deeper models [138]. The fundamental units for our ResNet architecture were 2 types of residual blocks, namely, basic and bottleneck blocks. A basic (or bottleneck) block consist of 2 (or 3) 1D convolutional layers, batch

normalization and ReLU non-linearity; the last layer's activation function was used following the addition of the learnt residual mapping with the input. Note that the number of filters, corresponding filter length and stride were tuned hyperparameters for each convolutional layer. Similar to 4.2.3.1, we also experimented with using a sine-cosine positional encoding to augment order information to our input. Figure 3.4 shows a sample ResNet architecture (top right) along with the design for basic (top left) and bottleneck (bottom left) residual blocks.

- Multi-scale residual neural network (MSResNet): Often, utilizing a fixed single-scale convolutional kernel size to extract features from only one scale may not be optimal. Consequently, we experimented with the multi-scale kernel-based ResNet architecture [139] to derive features from multiple scales. The extracted features from the initial convolutional block were sent through 2 branches of 3 basic blocks with [64, 128, 256] filters, where convolutional kernels in the 2 branches were fixed to be 3 and 5 (resp.). Next, these CNN-extracted multi-scale features were concatenated to a single vector. This vector was fed as input to a dense network with 2 output neurons (one for each class: HC and PwMS).

- Temporal convolutional network (TCN): TCN [140] utilizes residual connections as well as dilated causal convolutions, where dilations enable the model to look quite far back in the past while making predictions and causality ensures no future data leaks to the past. Note that each TCN block consists of a weight normalization layer (see [143] for details).

We further compared our multi-stride sequence classification with the 3 recurrent DL models described below.

- Vanilla recurrent neural network (RNN): RNNs intrinsically integrate the sequential order of strides as internal memory in their backbone architecture; this recurrence mechanism is not present in generic convolutional models. For recurrent layers, the output from the $(t-1)$-th time step is fed back into the network along with the input at t-th step to determine step t's outcome. The output features from the last RNN layer at the last time step are provided to multiple fully connected layers to output the class prediction probabilities.
- Long short-term memory (LSTM): LSTM [144] resolves the vanishing gradient problem that is existent in vanilla RNNs when dealing with longer sequences, given its feedback loop structure. A single LSTM unit, as depicted in Figure 3.6, utilizes a cell state and input, forget and output gates, to either include or eliminate data to the cell state. Similar to 4.2.3.1, our LSTM model consisted of a stack of n (hyperparameter) LSTM layers, where

each layer i produced a series of hidden size si (hyperparameter) number of features. We experimented with both uni- and bi-directional LSTM layers. The extracted features from the n-th LSTM layer at the last time step were followed with a fully connected network to yield the two output probabilities.

• Gated recurrent unit (GRU): GRU [145] also utilizes 2 gates, namely, reset and update gates to handle the vanishing gradient problem in recurrent networks. Our GRU model was a stack of n (hyperparameter) uni- or bi-directional GRU layers, where each layer i outputted a sequence of hidden size si (hyperparameter) features. The features from the n-th layer at the last time step were followed with a dense network to output class probabilities.

4.2.3.2 Modeltraining To prevent information leakage, we ensured that no single subject had its multi-stride sequences
split between training and validation folds. All computations were implemented on an NVIDIA GPU (12GB Tesla P100) using PyTorch v1.7.0 DL platform in Python 3.6. In all classifiers, we set a fixed random seed for reproducible results. We processed our data in batches of 128 samples each and randomly shuffled training samples at every epoch to reduce bias. We tried several optimization algorithms, namely, stochastic gradient descent with and without momentum, root mean square propagation (RMSprop), adaptive moment estimation (Adam), and Adam with decoupled weight decay (AdamW), each with different learning rate schedules as well as weight decay [165]. In addition to weight decay and early stopping (with patience (hyperparameter) epochs), using dropout between network layers also helped prevent over-fitting in our models. To manage the possible disparity in scales of the processed model features, we tried layer normalization [148] to normalize each feature to zero mean and unit variance. A thorough experimental hyperparameter search was performed on the validation set to determine optimal framework for each learning classifier.

4.2.3.3 Evaluationdetails

For evaluating our task generalization classifiers, we used the test set metrics, namely, precision (P), recall (R), accuracy (A), F1 score (F1) and area under receiver operating characteristic curve (AUC); for subject generalization, we used the mean and standard deviation in cross-validation metrics. All models were evaluated at 2 categorizations, namely, 5-stride sequence- and subject-level; majority voting was used to classify subjects as HC or PwMS. Thus a correctly classified

subject's walk had a majority of multi-stride sequences accurately detected as of the appropriate cohort. We denote the sequence and subject-level evaluation metrics with seq (i.e. Pseq, Rseq, Aseq, F1seq, AUCseq) and sub (i.e. Psub, Rsub, Asub, F1sub, AUCsub) in the sub script (resp.). Further, for all DL models, we monitored learning curves for convergence of training accuracy and cross entropy loss metrics across epochs.

4.3 Results

In general, MS subjects had a broader stride width and a shortened stride length; and additionally, a reduced cadence, speed and single support time along with a prolonged double support, stance and stride times. These observations are indeed aligned with the past findings regarding gait changes in PwMS. However, no single feature demonstrated a clear distinction between PwMS and HC; and thereby, a supervised learning approach is meaningful for this domain.

4.3.1 Prediction Models

In order to classify sequences and subjects between HC and PwMS for task- (4.3.1.1) and sub- ject generalization (4.3.1.2) designs, 16 diverse traditional machine and DL algorithms were compared with size-N and regress-N data. These sequences were fairly balanced across both classes in our data.

4.3.1.1 Taskgeneralization

Table 4.2 summarizes the sequence- and subject-wise evaluation metrics for the top-3 ML and DL task generalization classifiers on categorizing the test set sequences of trial WT. Majority voting evidently upgraded all the sequence-wise performance metrics within each algorithm, such as from 82.3% to 88.6% accuracy on MSResNet with size-N data. The results further improved across all metrics for DL algorithms with regress-N data in contrast to when using size-N data. The top-3 DL algorithms, that is, MSResNet, GRU and RNN (in order), had a sequence-wise accuracy, Aseq, of 91.7%, 88% and 87.3% (resp.), with the regression normalized data. Further, majority voting gave a perfect subject-level classification accuracy, Asub, across the 3 top DL algorithms. In contrast, these DL models had an Aseq of less than 85% and the maximum Asub of 97.1% when using the size-N data. Similarly, the highest sequence classification AUC (AUCseq) was 0.97 using MSResNet, followed by 0.95 and 0.94 using GRU and RNN (resp.), with the regress-N data, while the maximal AUCseq was 0.92 with size-N data using the vanilla

Table 4.2: Task generalization: comparing sequence- and subject-wise test set performance across top-3 ML and DL algorithms

			Sequence-based evaluation metrics					Subject-based evaluation metrics				
	Algorithm	Normalization	Accuracy	Precision	Recall	F1score	AUC	Accuracy	Precision	F1score	Recall	AUC
ML	DT	Size-N	0.731	0.698	0.661	0.679	0.759	0.886	0.882	0.882	0.882	0.918
		Regress-N	0.829	0.824	0.824	0.824	0.908					
	XGBoost	Size-N	0.751	0.700	0.714	0.719	0.755	0.886	0.933	0.875	0.824	0.978
		Regress-N	0.886	1.0	0.867	0.765	0.964					
	MLP	Size-N	0.799	0.796	0.757	0.753	0.895	0.914	0.938	0.909	0.882	0.930
		Regress-N	0.914	0.938	0.909	0.882	0.944					
DL	MSResNet	Size-N	0.823	0.842	0.711	0.779	0.916	0.886	1.0	0.867		0.943
		Regress-N	0.911	0.877	0.937	0.906	0.825	1.0	1.0	1.0	1.0	1.0
	RNN	Size-N	0.777	0.739	0.747	0.746	0.845	0.971	1.0	0.941	0.970	0.949
		Regress-N	0.844	0.844	0.781	0.813	0.887	1.0	1.0	1.0	1.0	1.0
	GRU	Size-N	0.823	0.802	0.797	0.791	0.897	0.971	1.0	0.941	0.970	0.949
		Regress-N	0.871	0.831	0.884	0.857	0.971	1.0	1.0	1.0	1.0	1.0
			0.821	0.851	0.754	0.800	0.915					
			0.840	0.857	0.811	0.834	0.913					
			0.880	0.856	0.867	0.861	0.971					
			0.915	0.915	0.915	0.915	0.971					
			0.913	0.915	0.913	0.913	0.944					

Note: ML is machine learning, DL is deep learning, AUC is area under the receiver operating curve, Size-N is body size-based normalization, and Regress-N is multiple regression-based normalization. The numbers in bold represent the highest model performance. See Section 4.2.3.1 for details on algorithms.

RNN architecture. The top-3 ML models, namely, DT, XGBoost and MLP, all had an Asub of less than 92% vs. a perfect Asub with DL models using regress-N data. MSResNet with regress- N data was an overall top-performer for task generalization with an accuracy, F1 and AUC of 91.7%, 0.91 and 0.97 (resp.), at a sequence-level, followed by GRU and RNN with a matching perfect subject-level classification. This top MSResNet architecture, illustrated in Figure 3.4, was trained for 45 epochs (as decided by the early stopping paradigm with patience 20) with a batch size of 100, and Adam optimizer along with a learning rate of 0.005; with nearly 2.1 million model parameters, this model took 15 minutes to train and 1.5 seconds to evaluate on a GPU. MSResNet utilizes both multi-scaled and residual learning frameworks to discover robust dynamics in gait motion.

4.3.1.2 Subjectgeneralization

Table 4.3 illustrates the mean and standard deviation of 5-fold cross-validation evaluation metrics for the top-3 ML and DL subject generalization classifiers. Not surprisingly, the subject-wise metrics were superior to the sequence-wise performance measures. Overall, ResNet was the highest-performing classifier across all DL and traditional ML algorithms. All algorithms performed better with the regress-N data in contrast to the standard size-N data. The top subject generalization algorithm was ResNet with regress-N data attaining the mean accuracy, F1 and AUC of 72.8%, 0.63 and 0.70 (resp.), at sequence-level; and 82.9%, 0.81 and 0.81 (resp.), at

Table 4.3: Subject generalization: comparing sequence- and subject-wise mean cross-validation performance across top-3 ML and DL algorithms

			Sequence-based evaluation metrics					Subject-based evaluation metrics				
	Algorithm	Normalization	Accuracy	Precision	Recall	F1score	AUC	Accuracy	Precision	Recall	F1score	AUC
ML	LR	Size-N	0.569± 0.07	0.462± 0.28	0.414± 0.16	0.417± 0.20	0.565± 0.09	0.629± 0.17	0.473± 0.27	0.567± 0.33	0.511± 0.29	0.593± 0.21
		Regress-N	0.616± 0.03	0.517± 0.26	0.50± 0.14	0.483± 0.19	0.633± 0.07	0.686± 0.11	0.513± 0.27	0.627± 0.40	0.550± 0.31	0.70± 0.13
	LSVM	Size-N	0.559± 0.06	0.461± 0.28	0.416± 0.15	0.413± 0.19	0.542± 0.10	0.60± 0.17	0.473± 0.27	0.533± 0.34	0.491± 0.30	0.593± 0.18
		Regress-N	0.619± 0.07	0.505± 0.29	0.508± 0.22	0.478± 0.23	0.641± 0.09	0.714± 0.13	0.553± 0.32	0.633± 0.40	0.572± 0.33	0.660± 0.21
	DT	Size-N	0.510± 0.05	0.419± 0.29	0.425± 0.27	0.382± 0.23	0.458± 0.09	0.543± 0.06	0.50± 0.32	0.517± 0.34	0.461± 0.25	0.563± 0.22
		Regress-N	0.642± 0.12	0.572± 0.19	0.583± 0.14	0.550± 0.13	0.578± 0.12	0.743± 0.17	0.713± 0.25	0.810± 0.26	0.721± 0.21	0.733± 0.22
DL	ResNet	Size-N	0.673± 0.05	0.545± 0.25	0.383± 0.24	0.428± 0.24	0.528± 0.21	0.743± 0.11	0.547± 0.15	0.850± 0.20	0.644± 0.14	0.693± 0.14
		Regress-N	0.728 ± 005	0.590± 0.23	0709 ± 026	0630 ± 022	0.70± 0.20	0829 ± 011	0833 ± 021	0810 ± 019	0808 ± 017	0813 ± 016
	TCN	Size-N	0.592± 0.09	0.503± 0.31	0.405± 0.17	0.423± 0.19	0.536± 0.17	0.629± 0.11	0.533± 0.09	0.633± 0.22	0.563± 0.13	0.647± 0.17
		Regress-N	0.670± 0.10	0606 ± 034	0.464± 0.23	0.493± 0.24	0.612± 0.15	0.743± 0.11	0.613± 0.21	0.767± 0.23	0.656± 0.17	0.640± 0.10
	RNN	Size-N	0.663± 0.06	0.561± 0.32	0.504± 0.24	0.505± 0.23	0.578± 0.19	0.714± 0.09	0.680± 0.19	0.677± 0.23	0.661± 0.18	0.747± 0.15
		Regress-N	0.671± 0.09	0.577± 0.18	0.416± 0.17	0.475± 0.17	0.635± 0.15	0.771± 0.07	0.553± 0.30	0.720± 0.37	0.613± 0.31	0.60± 0.08

Note: The numbers in bold represent the highest model performance.

subject-level classification. However, the top-3 ML models, namely, LR, LSVM and DT, all ended up with a mean Asub, F1sub and AUCsub of less than 75%, 0.73 and 0.74 (resp.). The top-performing ResNet architecture used a positional encoding layer followed by an initial convolutional block and 4 bottleneck blocks with {64, 128, 256, 512} filters (resp.); each bottleneck residual block had 3 convolutional layers with kernel sizes {3, 5, 1} (resp.). It was trained for 5 epochs with a batch size of 128 and AdamW optimizer (learning rate: 0.01, weight decay: 0.01); with nearly 433K trainable parameters, training took around 1.5 minutes on GPU. AdamW typically helps models to train faster and generalize better.

This diverse performance across model frameworks and normalization schemes by different classifiers is attributed to the no free lunch theorem for supervised algorithms, stating there is no universally accurate algorithm across all datasets and evaluation metrics [166]. For further post hoc analysis in sections 4.3.2 to 4.3.4, we used regress-N data and top-performing DL algorithms as they revealed better performance over both task and subject generalization frameworks.

4.3.2 Ablation Study on Gait Features

Next, we perform an ablation study in an attempt to comparatively evaluate the importance of various feature subcategories for classification, particularly 7 temporal, 4 spatial, 13 spatiotemporal, 8 kinetic, 15 temporal-kinetic and 12 spatial-kinetic features relative to all 21 features. All subject-wise evaluation metrics for the top model per data subset are presented in Table 4.4 across both the task- and subject generalization schemes. Note that we train and tune all ML and DL architectures entirely from scratch on these feature subsets as part of our post hoc analysis. DL, especially all recurrent architectures along with ResNet and MSResNet, exceeded the perfor-

Table 4.4: Ablation study for task and subject generalization frameworks

	Taskgeneralization						Subject generalization					
Feature set	TopDL algorithm	Asub	P_{sub}	R_{sub}	F_{1sub}	AUC_{sub}	TopDL algorithm	Asub	P_{sub}	R_{sub}	F_{1sub}	AUC_{sub}
T	ResNet	0.83	0.79	0.88	0.83	0.87	ResNet	071016.	047020.	090 ± 020	058019.	065013.
S	ResNet	0.83	0.87	0.76	0.81	0.90	ResNet	066015.	034031.	049042.	039033.	058015.
ST	GRU	0.97	0.94	1.0	0.97	1.0	ResNet	080007.	063024.	087019.	069018.	077008.
KIN	MSResNet	0.86	1.0	0.71	0.83	0.86	RNN	077015.	056046.	050045.	051043.	069025.
SKIN	MSResNet	0.91	0.94	0.88	0.91	0.90	ResNet	074011.	065013.	073027.	068019.	068013.
ALL		0.94	1.	0.88	0.94	0.97	ResNet	.077011	.065018	082019.	.071014	.074012
		1.0	0	1.0	1.0	1.0		083 ± 011	083 ± 021	081019.	081 ± 017	081 ± 016

Note: The numbers in bold represent the highest model performance.

mance of traditional ML algorithms over all feature subsets and both generalization frameworks. It is interesting to note that there is an overlap between top architectures in 4.3.1 and Table 4.4. The optimal task generalization metrics were observed by utilizing the entire 21-dimensional feature set with MSResNet (Asub: 1.0, F1sub: 1.0), closely followed by spatiotemporal subset with LSTM (Asub: 0.97, F1sub: 0.97), and then by spatial-kinetic parameters again with MSResNet (Asub: 0.94, F1sub: 0.94). Analogously for subject generalization, employing all features with ResNet resulted in top mean cross-validation performance (Asub: 0.83, F1sub: 0.81), followed by spatiotemporal (Asub: 0.80, F1sub: 0.69) and spatial-kinetic (Asub: 0.77, F1sub: 0.71) feature subsets, both also with ResNet. Across all feature designs for both model frameworks, CNN and TCN were never the top performers. Note that task generalization had a distinct variety of models that were top performers for each feature stream, whereas ResNet was the only top performer for subject generalization across all subsets, except for temporal-kinetic features where RNN was better. Within both task- and subject generalization designs, results when utilizing the complete feature set surpassed all other examined subset combinations; moreover, we observed that the spatiotemporal subset presents superior performance to any other composition with kinetic features. Further, comparing these results in Table 4.4 with the performance in [72], we infer that using DL with multiple strides outweighs single-stride performance across both model designs and all data subgroups. Overall, this analysis backs our use of all spatiotemporal and kinetic features for MS classification.

4.3.3 Feature Importance

In this section, we attempted to demonstrate the interpretability of our DL models by means of 1) permutation feature importance (4.3.3.1) that defined the most and least informative features in our gait classification models, and 2) visualizing the neural network's inner feature maps at

the penultimate layer (4.3.3.2) that gave insights about model's complex internal processing.

4.3.3.1 Permutationfeatureimportance Having fixed the regress-N normalization scheme, we permuted each of the 21 gait features,
one at a time, and assessed the reduction in evaluation metrics for our top performing models, i.e., MSResNet for task generalization and ResNet for subject generalization. The inherent randomness in shuffling might bias our findings, thus this procedure was repeated 20 times for the test set and the corresponding metrics over these reiterations were averaged out for relatively robust results. This shuffling procedure broke the relationship between the shuffled feature and the corresponding target, and thus the drop in model performance after feature permutation was indicative of how much the model depended on the feature. Therefore, a reduced model performance after permuting a feature signified higher dependence of our model on the associated feature for classification and consequently, a greater importance of the respective feature. For task generalization, force at MidSSR followed by right single support and right foot progression angle (in order) were the most informative features; however, walk ratio was the least predictive of labels. For subject generalization, the most informative features with the least accuracy after permutation were stride time and lateral shift followed by stride length. A few features, namely, cadence and lateral deviation, had very little effect on model performance for subject generalization.

4.3.3.2 Visualizingpenultimatelayerfeaturevectors Given a DL model's involved architecture comprising of several layers with numerous neurons
and correspondingly large number of learnt parameters, we visualize our top DL network's feature vectors at the penultimate step, in an attempt to comprehend its complex processing to some extent. Considering high dimensional feature space at our model's last layer, we firstly used a non-linear dimensionality reduction technique, namely, t-distributed stochastic neighbor embedding (t-SNE) [167], to collapse the multidimensional feature maps into a 2D t-SNE embedding space and subsequently visualize it. Primarily, t-SNE is an iterative optimization algorithm to assign multidimensional data points to a lower dimensional space such that closer points in the higher dimensions still stay close and likewise, farther points remain distant in the reduced space as well. More formally, t-SNE minimizes the difference, as statistically measured by the Kullback-Leibler (KL) divergence, between the two probability distributions measuring the pairwise similarities of the original data points and the corresponding points in the low-dimensional t-SNE embedding (resp.). Let $x_j \in \mathbb{R}^d$ denote points in the original d-dimensional space and

$y_j \in R2$ be the corresponding points in the 2D mapped space, then the original and reduced data similarity matrices are given by [pjk] (eq. (4.1)) and [qjk] (eq. (4.2)) (resp.).

$$p_{jk} = \frac{\exp(-||x_j - x_k||^2/2\sigma^2)}{\sum_{m \neq n} \exp(-||x_m - x_n||^2/2\sigma^2)} \tag{4.1}$$

$$q_{jk} = \frac{(1 + ||y_j - y_k||^2)^{-1}}{\sum_{m} \sum_{n \neq m} (1 + ||y_m - y_n||^2)^{-1}} \tag{4.2}$$

Note that we assume a Gaussian distribution around x_j with variance σ^2, and a t-Student distribution with one degree of freedom around y_j. Then, t-SNE minimizes eq. (4.3) via gradient descent to retain the original data structure while reducing dimensions.

$$KL(P||Q) = \sum_{j} \sum_{k \neq j} p_{jk} \log \frac{p_{jk}}{q_{jk}} \tag{4.3}$$

Our ultimate objective is to visualize any natural clusters of data points that may emerge in the penultimate space. Figure 4.4 visualizes the 2D t-SNE of the 512-dimensional last layer embedding for the best task generalization model (MSResNet). Clearly, two inherent clusters, green and red, categorizing the HC and PwMS 5-stride sequences (resp.), originate in the penultimate layer; further, there appears to be a progression among sequences of mild, mild-to-moderate and moderate severity subgroups within PwMS. These native arrangements demonstrate the robustness of our predictions and in fact validate that our feature space is well optimized by backpropagation to classify sequences in HC and PwMS. In conclusion, these results verify that our model extracted necessary information from the data that enabled t-SNE to clearly identify two distinct classes of sequences.

4.3.4 Association of Predictions with Lower Extremity Function

We examined a possible correspondence between the lower extremity physical function in middle-aged to older adults and our top DL model's prediction probabilities. We used the short physical performance battery (SPPB) assessment [150] to measure the physical performance of middle-aged to older adults on a scale of 0 (worst) to 12 (best). SPPB examined 3 areas that emulate day-to-day tasks essential for independent living, namely, static balance, gait speed and getting in and out of a chair. A higher summary score on SPPB signified none to mild mobility limita- tions and a lower score implied severe limitations. Our dataset had subjects with frailty ranging from minimal to moderate, i.e. SPPB: 10.37 ± 1.85 [6 – 12]. Figure 4.5 depicts the predictions made by the top-performing subject generalization model, ResNet, w.r.t the corresponding

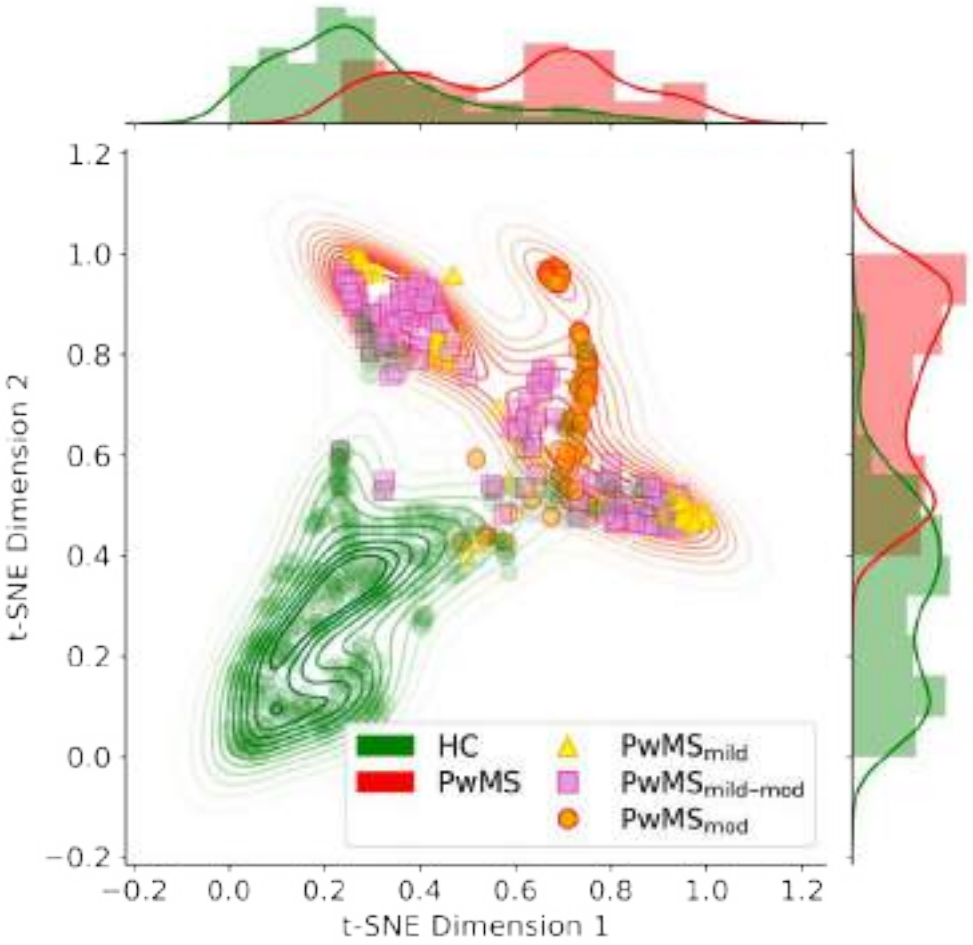

Figure 4.4: 2D t-SNE visualization for task generalization. Two natural clusters, shown in
green and red, grouping the 5-stride sequences of HC and PwMS (resp.), are identified in last layer embedding for the top task generalization model. Mild, mild-to-moderate and moderate severity subgroups within PwMS are marked in yellow triangles, pink squares and orange circles (resp.).

subject's SPPB. The markers, i.e. green-edged circles and red-edged squares, represent actual HC and PwMS (resp.), where marker face-color denotes the corresponding prediction probabil- ity for class HC. Horizontal axis displays the overall SPPB score and background stripe color depicts the predicted class by the model, i.e., markers on green and red stripes are predicted as HC and PwMS (resp.). Note that markers in each SPPB and prediction stripe are sorted in order of prediction probability of their true class. For instance, one true PwMS with SPPB of 10 was misclassified as healthy with a particularly high HC confidence probability and no true HCs were misclassified at SPPB 10. We observed from Figure 4.5 that PwMS with a higher SPPB, i.e. bet- ter physical performance, had majority of 5-stride sequences incorrectly predicted as belonging to the healthy cohort; this was quite likely as gait characteristics of PwMS with none to mild

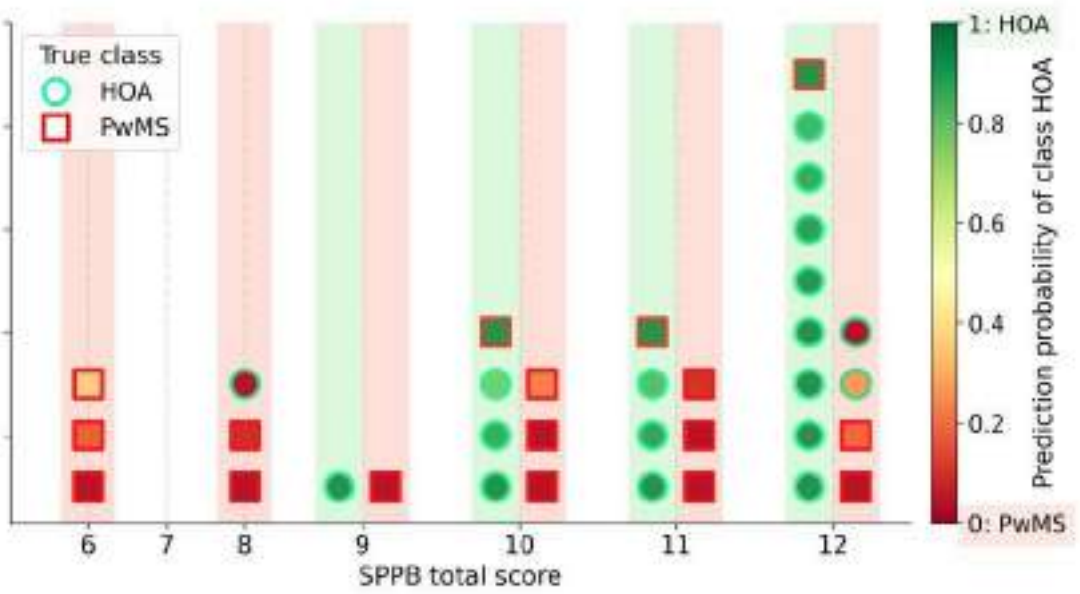

Figure 4.5: Visualizing the predictions of subject generalization model w.r.t the corresponding subject's lower extremity function. Green-edged circles and red-edged squares represent actual HC and PwMS (resp.), where marker's face color shade denotes the corresponding prediction probability for class HC. Horizontal axis displays the overall SPPB score and background stripe's color depicts the prediction surface, where markers on green (or red) stripe are predicted as HC (or PwMS) by the model.

mobility limitations could be difficult to discern from healthy gait. In summary, SPPB seemed to have some correspondence with our model predictions, but, no significant correlations were observed.

4.4 Discussion

We studied a DeepMS2G framework employing data driven DL on 21 multi-stride spatiotempo- ral and kinetic gait features to classify middle-aged to older adults with and without MS.

Our proposed system offered an automated, accurate and remote monitoring mechanism for neurological gait classification and was quicker, utilizing only 5-stride sequence or a brief gait length than most typical clinical gait assessments. Past works [113, 155] have explored DL with domain knowledge-based gait features for low vs. high fall risk assessment. However, compared to our exhaustive experimental comparison with 16 diverse models across 2 different designs, namely, task- and subject generalization, these past works have only ex- amined LSTM and 4 traditional ML classifiers for subject generalization. Further, these studies

focused on 4 kinematics-based gait parameters, whereas we utilized dynamics from 21 kinematic as well as kinetic features. Additionally, we comprehensively investigated the explainability of our top-performing algorithms via post hoc analysis (4.3.2 to 4.3.4), which was absent in pre- vious analogous studies. Although our prior research [72] using traditional ML frameworks on individual strides provided utility in the identification of MS-related changes in gait, our current approach employing DL with multi-stride data provided a tool to extract additional information from gait dynamics and variations across temporally ordered strides. Moreover, using DL and multi-stride dynamics for MS classification exceeded the subject-wise performance metrics pre- sented in [72] across both task- and subject generalization designs. A few additional past works have explored using classical ML to classify MS based on gait data [96, 97], however, to the best of our knowledge, ours is the first study extensively examining modern DL algorithms on multi-stride spatiotemporal and kinetic gait features for MS classification.

In contrast to prior work using wearable inertial measurement unit sensors [97, 113, 123, 168, 169] electromyography (EMG) [160, 170], and motion capture systems [161] to predict neuromuscular changes in neurological gait, our approach requires no sensors to be placed on the participant, which simplifies data collection. However, there is the need for an instrumented treadmill, which limits usability in smaller or rural medicine practices. Further, depth cameras capturing 3D movement patterns have been explored for gait assessment [129, 171], but these systems are relatively costlier, have some limitations when used outdoors and are constrained by the camera to object distance. Most past studies explored either an end-to-end DL framework that demanded larger datasets or a traditional ML approach on hand engineered features that more suited smaller datasets. We studied a hybrid approach utilizing our domain-knowledge based spatiotemporal and kinetic feature space with advanced DL methods in an effort to overcome the challenges of limited clinical data, which exist in most medical scenarios.

The advantages of using regression normalized gait features were apparent when regress-N improved the accuracy of identifying pathological gait than the standard size-N strategy in both task and subject generalization frameworks. ResNet-based models, namely MSResNet for task generalization and ResNet for subject generalization, were top-performers across both model designs, which might guide model selection in future studies on neurological gait classification. Also, using a 5-stride data sample allowed for frequent conclusions, which might assist in the deployment of future clinical applications based on this work. Moreover, our analysis in sections 4.3.2 to 4.3.4 helped establish the interpretability of our top DL algorithms, which might facilitate gait practitioners to comprehend and trust the findings from our proposed system. An ablation study on the set of features in 4.3.2 supported using all the extracted gait features for better predictability in both model task and subject generalization designs. Moreover, we observed that the spatiotemporal subset presents superior performance to any other composition with ki-

netic features. When only including a subset of features to examine the most relevant features driving the DL performance, we found that stride and single support times, force at midstance, and butterfly diagram-based lateral shift were the most valuable features across both classifica- tion frameworks. Further, we observed that PwMS with none to mild mobility limitations had the majority of 5-stride sequences incorrectly predicted as belonging to the healthy cohort. This is in line with observations in some past studies on MS, where gait parameters were noticed to worsen for severely affected MS patient groups compared to the control group and are not seen in PwMS with mild disability [172].

A larger study would allow better understanding of the dependence of the regression function on demographics, and also better understanding of confidence intervals. A broader sample of neurological disability levels in PwMS and subtypes might assist in making more generalized predictions for the heterogeneous MS community. Assessing gait in additional concurrent cognitive settings in future works might provide increased sensitivity to distinguish between MS related changes. Recently, transformer-based DL models have achieved outstanding performance on several vision and language tasks [173, 174]. Given higher model complexity in transform- ers and our relatively smaller dataset for classification, we did not consider transformer-based models for this work. Future work might involve evaluating the performance of transformers for neurological gait classification. Lastly, clinical applications of gait-based data to determine MS related changes might benefit from further examination using wearable gait-related data in real-world environments.

4.5 Summary

We proposed a DeepMS2G pipeline for classification of PwMS using DL and multi-stride dy- namics across domain knowledge-based spatiotemporal and kinetic gait features. We evaluated DeepMS2G to generalize over distinct walking trials and new participants. We observed that ResNet-based models with regression-based normalization were top performers across both task and subject generalization designs. With no known cure and clinically unpredictable disease progression, our proposed framework might augment findings from standard clinical tests and aid clinicians in defining effective medication strategies. Our ultimate objective is a system to be used in hospitals, homes or workplaces to automatically provide an automated diagnosis of movement dysfunctions in older adults.

CHAPTER 5

LEARNING AN ACCELEROMETER DATA-BASED FRAMEWORK

In this chapter, we review the work in Exploration of Machine Learning to Identify Community Dwelling Older Adults with Balance Dysfunction Using Short Duration Accelerometer Data.
This work previously appeared as [75].

5.1 Introduction

Every year, approximately 3 million people are hospitalized due to fall injuries leading to increases in health care costs [175]. There are several risk factors that can lead to falls such as poor eyesight, home hazards, walking and balance dysfunction [176]. Among these risk factors, balance dysfunction is considered one of the most common risk factors of a fall, which may provide an opportunity for early identification in community dwelling older adults through the use of instrumented tests. However, the cost to carry out objective and instrumented balance analysis in the clinical setting is fairly high, as it requires trained clinical personnel and the use of expensive equipment, such as a computerized dynamic posturography system.

The motor control test (MCT) assesses the reactive postural control reflex capacity of individ- uals by measuring the latency in the onset of a postural response after an anterior or posterior translation. The MCT provides a quantitative measurement that has been shown to be a sensitive indicator for the decline of postural control in community dwelling older adults [177]. Tradi- tional balance tests such as the Mini Balance Evaluation Systems test and Berg Balance test have also been used to predict balance dysfunction and fall risk, but these subjective measures take time and clinical professionals to administer the test.

Earlier studies have explored accelerometer-based predictive models and treadmill-based gait models to predict fall risk and neurological diseases in older adults using machine learning (ML) techniques [91, 95, 178]. The success of these ML techniques in the prediction of fall risk motivates the study of balance dysfunction prediction. In this study, we used MCT to differentiate community-dwelling older adults with either high or low dynamic balance ability for use in the training and testing of several notable ML algorithms that use accelerometer-based data from 2

sensors while walking. Current studies support that hip and knee motion are distinctive in differentiating balance ability during walking [179–181]. Thus, we aim to understand the features of accelerometer data collected from hip and knee that can be used to distinguish between persons who have balance dysfunction from those without.

In this study, we aim to validate the feasibility of wearable sensors to identify older adults who have trouble with balance. This technique may provide a low cost monitoring tool for patients who are on the edge of dynamic balance dysfunction.

5.2 Methods

The protocol for the study was approved by the Institutional Review Board under IRB number 17010. To record the gait patterns of participants, a 1-meter wide and 3-meter long Motek-ForceLink C-MILL instrumented treadmill [182] was setup to carry out an adaptive speed control trial, which allowed participants to walk at a self-selected comfortable pace. The experimental paradigm consisted of two tasks: Normal walking with a self-selected speed (NW) on the instrumented treadmill; and an MCT test (small, medium and large perturbations on force plate) on a Neurocom Clinical Research System. Two commercial accelerometer sensors (Delsys) were placed on the right hip and knee so as to record triaxial acceleration data at a 148.1 Hz sampling rate.

5.2.1 Study Participants

We recruited 21 participants from the local community to participate in this study. The Motor Control Test (MCT) was used to assess balance in subjects and categorize them as low and high balance participants. We chose 143.4ms as the cutoff MCT latency for low and high balance, which is the average MCT latency for community dwelling older adults [177]. Based on this benchmark, twelve participants were grouped into high balance (age: 66.75 (4.08) years; Tele- phone initial cognitive screening (TICS): 25 (4.16)) and nine participants were grouped into low balance (age: 66.44 (6.31) years; TICS: 25.89 (5.17)) respectively.

Table 5.1: Basic characteristics for the MCT and raw Hip (H) and Knee (K) accelerometer data

Groups	All subjects	High Balance	Low Balance	p-value
N	21	12	9	NA
MCT	140.24 ± 11.36	131.25 ± 4.99	152.22 ± 3.42	< 0001 *
Hx	2.05 ± 0.47	2.19 ± 0.57	1.88 ± 0.15	019
Hy	2.15 ± 0.94	2.35 ± 1.21	1.88 ± 0.15	072
Hz	1.76 ± 1.34	1.90 ± 1.58	1.58 ± 0.95	091
Kx	1.75 ± 0.40	1.85 ± 0.31	1.62 ± 0.47	016
Ky Kz	2.00 ± 1.29	2.21 ± 1.66	1.70 ± 0.33	066
	2.43 ± 1.74	2.67 ± 1.95	2.11 ± 1.41	019

5.2.2 Data Analysis

In order to attenuate noise in the signals, accelerometer data was filtered using a band pass butterworth filter with a 0.1 Hz high-pass and 30 Hz low pass cutoffs (eq. 7.3).

$$H(_{j\omega}) = \frac{1}{\sqrt{1 + \epsilon 2(\frac{w}{wp})^{8}}} \tag{5.1}$$

Two independent walking trials were collected from 21 subjects, with up to 120 seconds of continuous walking data collected in each trial. For this analysis, independent 60 second segments of accelerometer data while walking were used, where available. Overall, this resulted in a total of 82 data segments that were labeled based on the corresponding participant's MCT score. Figure 5.1 compares 10 seconds of x, y, z-axis acceleration from the two accelerometers at the hip and knee for a representative subject with low or high balance function. The acceleration data from the two accelerometers were collected in raw mode and were transformed into actual g-units. Table 5.1 report the peak frequency of the raw data in six dimensions. There is no statistical difference between the raw data peak frequency. For each time point, the resultant acceleration (ri) was calculated as follows:

$$r_{\overline{i}} \quad \sqrt{x_{i}^{2}+2+y_{i}^{2}} \quad z_{i} \tag{5.2}$$

where x, i y_i z_i are the i^{th} measurement sample of the raw acceleration signal in x-, y- and z -directions.

Traditional characteristics of recorded gait data accelerations, such as the mean (μ_h, μ k), stan-

dard deviation (σh, σk), coefficient of variance (CVh, CVk), root mean square (RMSh, RMSk), auto correlation (Ah,Ak), mean amplitude deviation (eq. 5.3) [183] (MADh,MADk), sig- nal magnitude area (eq. 5.4) (SMAh,SMAk), signal energy (area under the squared magni- tude) (SE,SE),pairedcorrelationcoefficient(hhhkkkhkrxy,ryz,rxz,rxy,ryz,rxz),peak-to-peakmean (μP2P,μP2P),standarddeviation(P2PP2P)andmaximum(P2PP2Phkσh,σkyh,yk)wereexaminedin

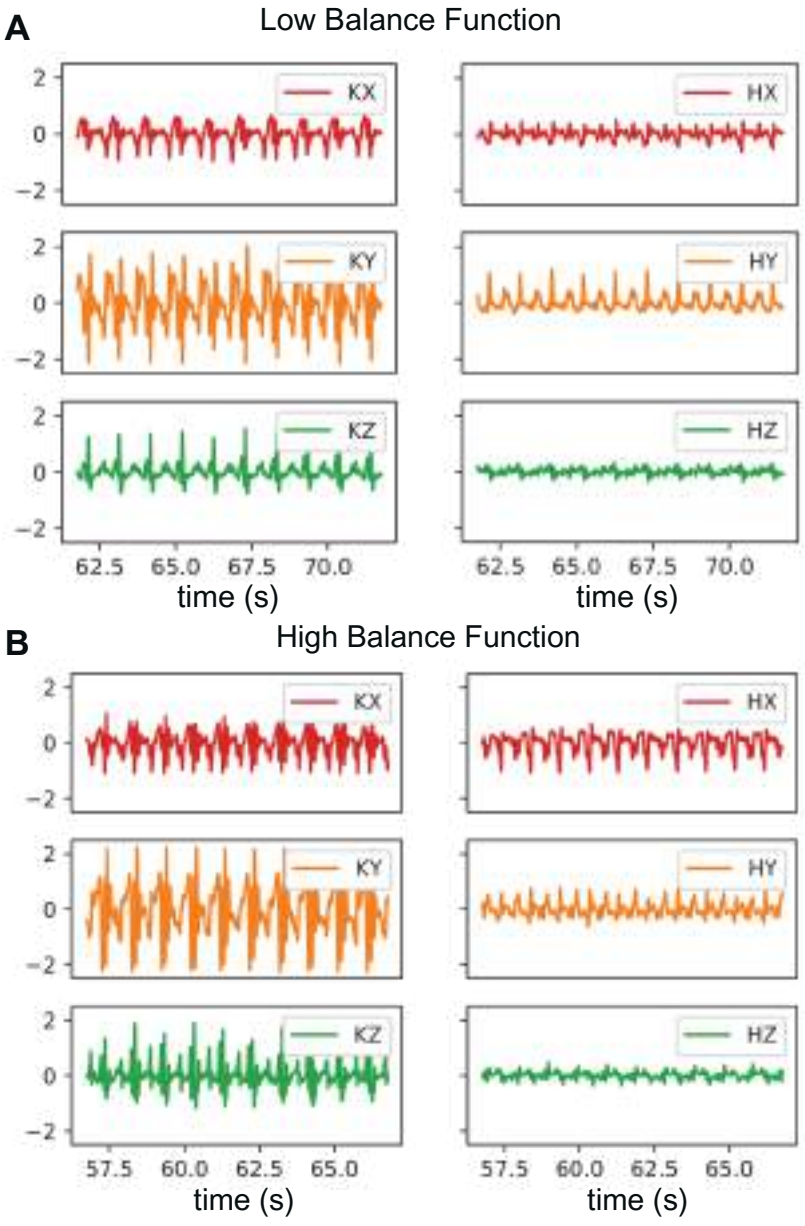

Figure 5.1: Filtered acceleration coordinates. Comparing 10 seconds of x, y, z-acceleration coordinates from the two accelerometers at hip (H) and knee (K) for a subject with low (A) and high (B) balance.

the accelerometer sensors placed on the knee and hip for differences in community dwelling older adults with low or high balance function. Figure 5.2 compares the distributions of ac- celerometer derived features between subjects with low and high balance function. No pattern distinguishing subjects with low or high balance function are evident in the exploratory plots. No statistically significant differences between subject groups in all features, were found from independent t-tests.

$$MAD = \frac{1}{n} \sum |r_i - \bar{r}| \tag{5.3}$$

$$SMA = \frac{1}{T} \int_0^T (|x(t) - \bar{x}| + |y(t) - \bar{y}| + |z(t) - \bar{z}|) dt \tag{5.4}$$

All features were normalized between 0 and 1 to avoid uneven ranges of features to dominate

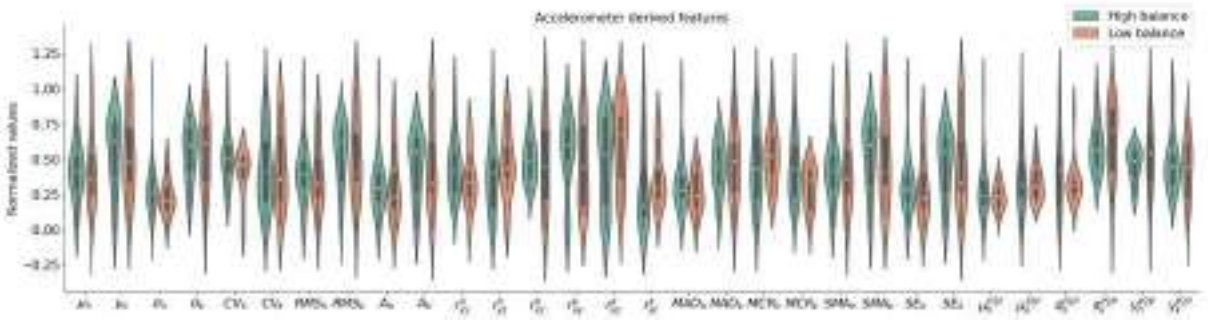

Figure 5.2: Accelerometer derived features. All features are scaled between 0 and 1 to plot on the same axis. Features are labelled by their corresponding mathematical notations.

the performance of ML models. The performance of seven state-of-the-art supervised learning classifiers, i.e. decision tree, ensemble random forest, support vector machine (SVM) with linear and radial basis function (RBF) kernel, gradient boosting machine (GBM), adaptive boosting (Adaboost) and eXtreme gradient boosting (XGBoost) were compared. Since our data set was limited in size, 5-fold cross-validation was used, where algorithms were repeatedly trained on randomly selected four parts with 60 seconds signals from sampled 32 trials and validated on the fifth part with 60 seconds signals from the 8 remaining trials each time until all 5 subsets were utilized in validation. We evaluated the performance of the models on the basis of accuracy,

F1-score and area under the curve (AUC) of the receiver operating characteristic (ROC) curve.
The implementation of analysis was done using Python via open source libraries such as sklearn, pandas, and scipy.

5.3 Results

Table 5.2 compares the cross-validation performance of the ML algorithms and reports the optimal performance metrics and hyperparameters. In classification of balance using the 60 seconds

Table 5.2: Mean cross-validation performance and tuned optimal hyperparameters for the ML algorithms

Algorithm	Accuracy	F_1 score	AUC
DT	0780 ± 016	0709 ± 024	0782 ± 016
RF	0890 ± 009	0860 ± 012	0865 ± 011
LSVM	0646 ± 012	060 ± 006	0705 ± 024
RBF SVM	0915 ± 007	0889 ± 009	0957 ± 008
Adaboost	0866 ± 012	0858 ± 012	0875 ± 014
GBM	0915 ± 009	0893 ± 011	0965 ± 005
XGBoost	0902 ± 013	0899 ± 012	0893 ± 017

accelerometer data, GBM algorithm provided the best average cross-validation performance with 91.5% accuracy, F1 scores of 0.90 and AUC of 0.97, followed by RBF kernel SVM by a slight margin with an AUC of 0.96. Figure 5.3 shows the ROC curve for each of the algorithms in this study. As the discrimination threshold varies, GBM and RBF SVM algorithms shows consistent low false positive rates and high true positive rates.

5.4 Discussion

This study uses wearable sensors while walking to classify balance dysfunction, a substantial contributor to fall risk, using ML in community dwelling older women. Consistent with recent studies examining predictors of fall risk using accelerometer data in older adults [184–187], the results of this study suggests that accelerometer data can be used to classify community dwelling older adults with balance dysfunction. With a 91% accuracy, a GBM algorithm, together with data from a wearable accelerometer, shows much promise at replacing expensive clinical and laboratory equipment for assessing the potential risk of dynamic balance dysfunction. In contrast with past ML approaches using accelerometer data to classify older adults with fall risk, which have used a random forest algorithm [188], [184], we found GBM to provide improvements in

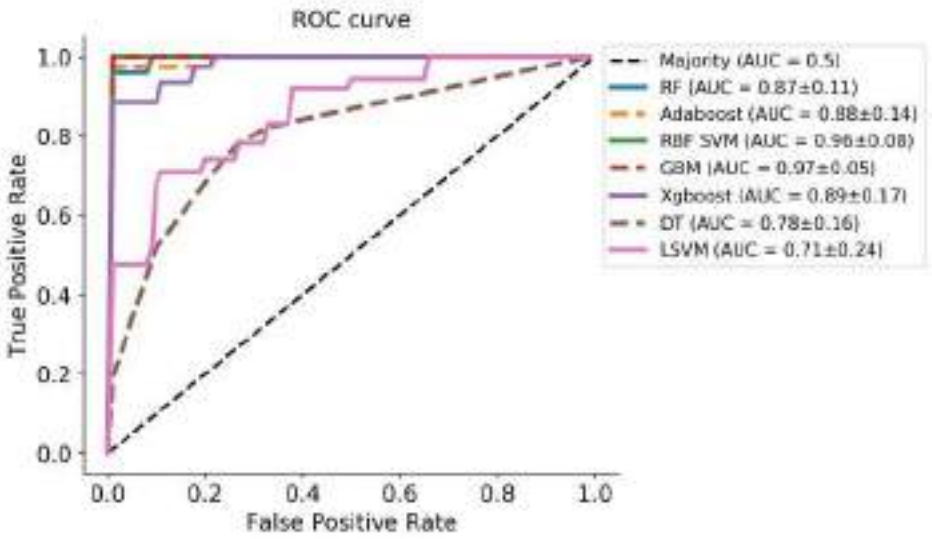

Figure 5.3: ROC curve. AUC of 0.5 means random guess and 1 signifies perfect classification.

overall classification performance.

The two sensor systems that were implemented in this study may provide a potential opportu- nity for the tracking of additional lower extremity movements, in comparison to prior one sensor systems [189]. However, additional work remains in examining changes in performance with 1 or 2 sensor systems.

The data used in this study demonstrates the feasibility of wearable sensor data to predict balance dysfunction. Compared to the instrumented balance tests such as the MCT and the instrumented time up and go test, this technique is inexpensive and can be used in real time. Also, the instrumented time up and go test is limited by the need of a clinical professional to administer the test in a controlled setting [190]. On the other hand, accelerometer based ML techniques can be used by a non-clinical professional in any setting.

5.5 Summary

The incidence of fall-related injuries in older adults is high. Given the significant and adverse outcomes that arise from injurious falls in older adults, it is of the utmost importance to iden- tify older adults at greater risk for falls as early as possible. Given that balance dysfunction provides a significant risk factor for falls, an automated and objective identification of balance dysfunction in community dwelling older adults using wearable sensor data when walking may be beneficial. In this study, we examined the feasibility of using wearable sensors, when walk- ing, to identify older adults who have trouble with balance at an early stage using state-of-the-art

machine learning techniques. We recruited 21 community dwelling older women. The experimental paradigm consisted of two tasks: Normal walking with a self-selected comfortable speed on an instrumented treadmill and a test of reflexive postural response, using the motor control test (MCT). Based on the MCT, identification of older women with low or high balance function was performed. Using short duration accelerometer data from sensors placed on the knee and hip while walking, supervised machine learning was carried out to classify subjects with low and high balance function. Using a Gradient Boosting Machine (GBM) algorithm, we classified balance function in older adults using 60 seconds of accelerometer data with an average cross validation accuracy of 91.5% and area under the receiver operating characteristic curve (AUC) of 0.97. Early diagnosis of balance dysfunction in community dwelling older adults through the use of user friendly and inexpensive wearable sensors may help in reducing future fall risk in older adults through earlier interventions and treatments, and thereby significantly reduce associated healthcare costs.

Part II

Clinical data analysis for prediction of disease progression

CHAPTER 6

LEARNING INTERACTIVE MODELS TO PREDICT DISEASE PROGRESSION TRAJECTORY AND CLINICAL SUBTYPES

In this chapter, we review the work in Interactive models for predicting Alzheimer's disease progression trajectory and clinical subtypes using machine learning [1] This work is currently in review as [76], and an earlier version previously appeared in the Machine Learning for Health workshop at the conference on Neural Information Processing Systems (NeurIPS) 2018, as [77]. The code for this work is entirely open source, and freely accessible at
for interested researchers.

6.1 Introduction

Alzheimer's disease (AD) is a progressive, age-associated neurodegenerative disease affecting memory, intellectual ability, and other cognitive domains. It is the most common form of dementia, affecting about 47 million people worldwide. AD is a clinically heterogeneous condi- tion, showing variation in symptom manifestation and disease progression rates. After the age of 65, the prevalence of dementia doubles every five years and increases exponentially after 90 years [191]. It poses a severe and ever increasing socioeconomic challenge [192].

AD pathological changes occur 20 years or earlier before the actual disease symptoms manifest [193–198]. In the absence of a cure or disease-modifying therapy, current management strategies are limited to supportive measures and are modestly effective symptomatic treatment [199, 200]. A major challenge for AD prediction is the presence of inherent phenotypic diversity in the AD population, limiting diagnosis, prognosis, and counseling of affected patients regarding their individual risks and expected progression rate. This problem becomes particularly burdensome as we move increasingly toward early-stage clinical trials when therapeutic interventions are likely to be most effective. The ability to predict and account for even a proportion of the disease course can significantly reduce the cost of clinical trials and increase the ability of such trials to detect treatment effects. Therefore, defining distinct disease subtypes and

[1] This research was done in collaboration with Vipul Satone, Anant Dadu, and Faraz Faghri.

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 notable downstream damage occurs [201]. Patients diagnosed with amnestic mild cognitive im- pairment (MCI) are at a higher risk for progression to dementia, but not all patients with MCI end up developing AD [202]. Research has been done to detect AD in patients with MCI or predict the early stage of AD using cerebrospinal fluid (CSF) biomarkers [203, 204], while oth- ers [205] have used psychometric and imaging data for predicting the progression of dementia in patients with amnestic MCI. In an implementation of a multiclass classifier using clinical and magnetic resonance (MR) brain images to classify controls, MCI, and AD patients, 79.8% ac- curacy was achieved [206]. Less research has been done on using only clinical data to predict the AD progression rate. Dadu et al. [207] used machine learning to classify Parkinson's disease patients into three different sub-categories with highly predictable progression rates. They ex- plored variations in onset and progression velocity and observed clusters of the motor, cognitive, and sleep disturbance related features using only clinical data. Here, we extend this approach by applying it to the clinical features of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. This study continues the investigation regarding early AD onset and progression started in Satone et al [77]. This study is designed to describe and predict the clinical progression of AD at an early stage after diagnosis.

6.1.1 Related Work

6.1.1.1 Evidencebeforethisstudy

In a systematic survey of literature surveying disease subtyping in Alzheimer's disease and re- lated dementias subtyping, we identified nearly fifty independent reports. Most of these reports separately looked at either progression or subtyping but not generally the two together. Addition- ally many of these papers focused on specific biomarkers, imaging metrics or clinical outcomes in a semi-hypothesis driven manner.

6.1.1.2 Addedvalueofthisstudy

The added value of this study is that it is a completely hypothesis-free, data-driven effort that incorporates current best practices in longitudinal machine learning (long short-term memory modeling) to predict peri-diagnostic trajectories in Alzheimer's disease with high accuracy in

an open science context. We identify and accurately predict three subtypes of Alzheimer's disease with varying rates of disease progression. In addition, we provide an interactive website for clinical researchers to predict the clinical subtype of an AD patient based on clinical parameters.

6.1.1.3 Implicationsofalltheavailableevidence We have identified a fast progressing subset of the Alzheimer's population that may be optimal for inclusion of clinical trials. Faster progressing cases may likely be more ideal for enrollment as they can potentially show successful drug readouts in a shorter period of time thus allowing for shorter, more efficient trials. We have also made these models easily available and deployable for public applications in drug development and so that other researchers can improve upon it over time as larger

6.2 Methods

6.2.1 Open Source Code and Data

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 ma- chine learning analysis) was made publicly available under GPLv3 as part of the supplementary information at The data used in this study was access-controlled from the Alzheimer's Disease Neuroimaging Initia- tive which requires individual sign-up to access the data. To aid researchers in using the tools presented in this manuscript as well as hopefully democratize complex machine learning practices for the clinical community, models can be deployed

6.2.2 Study Design

From a technical perspective, this study 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 AD, summarizing extensive 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 AD into three clinical subtypes, defined as slow, moderate , and fast progressors. Our analysis found that clinically related measures corresponding to memory and cognition more generally make up the AD progression space. Clinical data collected at baseline (study entry), after 6 and 12 months, were 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 pre- dict disease progression rates (control, slow, moderate, and fast) after 24 and 48 months from baseline. Further, we performed an in-depth analysis of the findings. We examined the reversion instances of AD captured in the constructed progression space, the correlation of Apolipopro- tein 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 6.4).

6.2.3 Participants

The ADNI study was launched in 2003 as a public-private partnership, led by Principal Investi- gator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial mag- netic 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. The study protocols for ADNI were approved by the Institutional Review Board. The ADNI study was carried out in phases, namely, ADNI 1 begin- ning in 2004, followed by ADNI GO in 2009, ADNI 2 in 2011, and ADNI 3 in 2016. These editions had different participants and data collection procedures, accounting for advancement in technologies. For more up-to-date information, ADNI 2 study is the main cohort in our study, and has the longest duration, with the greatest data availability, but also the greatest heterogeneity. The key eligibility criteria for the ADNI participants are highlighted in Section C.1.1 in the Appendix, and further details on the protocol can be found on the ADNI's procedure manual [208]. All participants went through comprehensive functional,

cognitive, and clinical assessments and provided a blood sample for APOE genotyping at their baseline visit (study entry). Additional details on the ADNI participants can be found in the Appendix Section C.1.2 (see Tables C.1 and C.2). We considered a total of 147 clinical variables (features) from the assessments mentioned in the Appendix Section C.1.3. An elaborated list of features used in each test with their definition is given in the Appendix Table C.3.

6.2.4 Procedures and Statistical Analysis

Only the observations which had data recorded for all the considered tests were taken into account. We used measurements taken at baseline and on visits after 6 and 12 months from the baseline to construct the AD progression space which was then extrapolated to 48 months post-enrollment in the study.

6.2.4.1 ADNIprogressionspaceandthepredictionmodel We leveraged 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. A descriptive plot with the number of observations available for each feature before and after data imputation is given in the Appendix Figure C.1. 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 multimodal. Furthermore, min-max normalization did not affect categorical features. Figure 6.1 shows the detailed workflow followed during the analysis. To reduce the dimensionality of the dataset, NMF [209] (with a rank of 2) was used on 582 observations with available data for baseline, and 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 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.

Next, unsupervised clustering via GMM [210] was used to define the hidden subtypes within

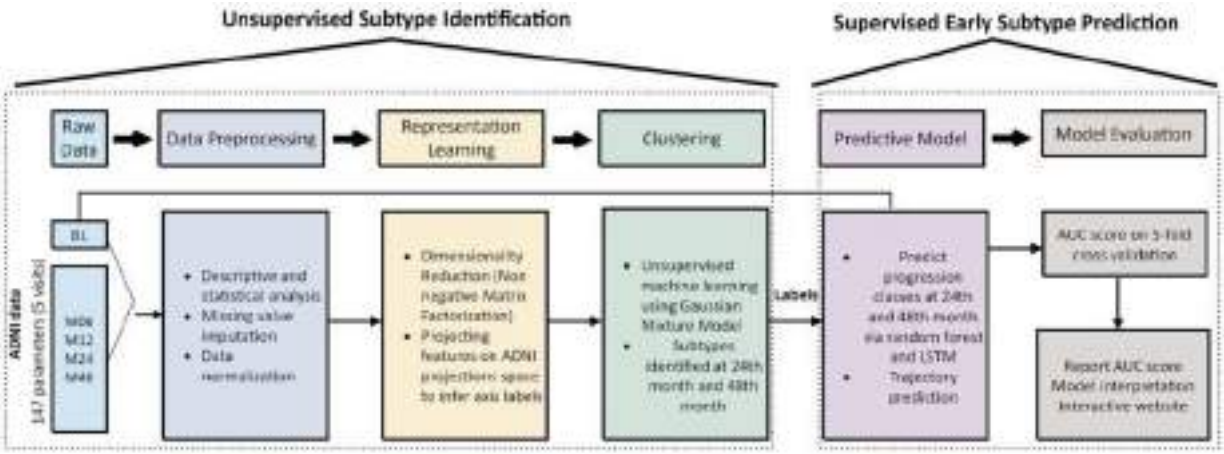

Figure 6.1: Workflow of analysis and model development.

the MCI and dementia patients. GMM is an expectation-maximization algorithm that maxi- mizes the likelihood of observing the data, given the underlying parameters of the distribution. Bayesian information criterion [211] was used to select the optimum number of underlying clus- ters for the GMM. Bayesian information criterion is a maximum likelihood estimate which tries to select the best model among the given set of candidates. After obtaining the AD progres- sion 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: predicts pro- gression at 24th month after baseline by using baseline and first-year factors b) Model 2: predicts progression at 48th month after baseline by using baseline and first-year factors.

Recurrent neural network (RNN) architecture with LSTM was also used to predict the pro- gression rates (slow, moderate, and fast) 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 probabil- ity of 0.2. Since our dataset size was limited in terms of the number of observations, five-fold cross-validation was used to evaluate the models. Among all the explored algorithms, a random forest classifier [212] gave the best five-fold cross-validation accuracy. Hence, parameters for the random forest algorithm were fine-tuned using grid search (repeated and summarized over 4800 iterations) and five-fold cross-validation.

6.2.4.2 Modelevaluation Sensitivity and specificity are measures of the proportion of true positives that are correctly
identified and true negatives, respectively. The plot of sensitivity on the y-axis and 1-specificity on the x-axis is called the area under the receiver operating characteristic (AUC of ROC) curve with a greater value representing a higher predictive performance. Since this is a multiclass problem, one versus all approach was used to calculate the AUC for each class. Next, five-fold cross-validation 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.

6.2.4.3 Trajectoryprediction

Having the constructed progression space and ADNI patient's progression trajectories in this space, we used the space and identified the trajectories to train a machine learning model capable of predicting prognosis for new patients. This model predicts the progression trajectory of patients at the 24th and 48th month. We used LSTM to train the model. After hyperparameter tuning using grid search, a bidirectional LSTM with 2 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 (Hochreiter and Schmidhuber 1997). Since the projection was done in the 2D axis, the mean squared Euclidean distance was used to assess the performance of this model. Only the features which were present for all the first three visits (baseline, 6-month, and 12-month visits) were considered for this study.

In the subsequent analysis, we studied the share of different frequencies of APOEε4 variants for each progression subtype since APOEε4 genotype is closely related to AD risk [213]. 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 genotype, and other selective critical features.

6.3 Results

We identified two progression indicator vectors and three clusters of AD patients relating to varying rates of progression. 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.

6.3.1 ADNI Progression Space

The observed axis labels for the features using 2D NMF and GMM are presented in the Appendix Table C.4. Progression indicator, i.e., coefficient matrix obtained from the NMF, was used to determine the hidden features that each of the two axes of the reduced space represents (as depicted in Figure C.2 in the Appendix). Progression indicator vectors represent latent features of the reduced progression space. Progression indicator coefficients for each feature are plotted in Figure 6.2, and they are separated by drawing a line with slope 1.

This transformed data was used to project the participant's disease progression stage at the 24th (Figure 6.3-a) and the 48th month (Figure 6.3-b). Further, three different zones, namely slow, moderate, and fast progression rates, were identified in the MCI and dementia patients at 24th and 48th month, as depicted in Figure 6.4.

6.3.2 Clinical Characteristics of the Identified Subtypes

To validate AD subtypes observed at 24 months from baseline, we compared the proportions of MCI and dementia subjects within each subtype (Table 6.1). As expected, the fast progression rate subtype consists of predominantly dementia patients (90.57%). Slow progression rate had no dementia patients and contained only MCI patients. The moderate progression rate subtype is dominated by MCI patients, which covered around 70% of the subjects in the moderate subtype. A similar trend was observed after 48 months from baseline as well (slow progressors have no dementia, moderate progressors have 5% dementia, and fast progressors cover about 72% of dementia patients).

APOE has three common alleles, ε2, ε3, and ε4, of which the ε4 allele is closely associated with increased risk of AD [213]. The distribution of APOEε4 alleles for each progression rate subtype is shown in Table 6.1, and Figure 6.5 after 24 and 48 months from the baseline, respec- tively. Figure 6.5 lists the proportion of APOEε4 variants in each of the identified subtypes. We observe that the progression rate increases with the number of APOEε4 alleles. See Section 6.3.3.2.

We also evaluated the distribution of memory and cognitive decline in identified subtypes. Details of this are shown in the Figure 6.12.

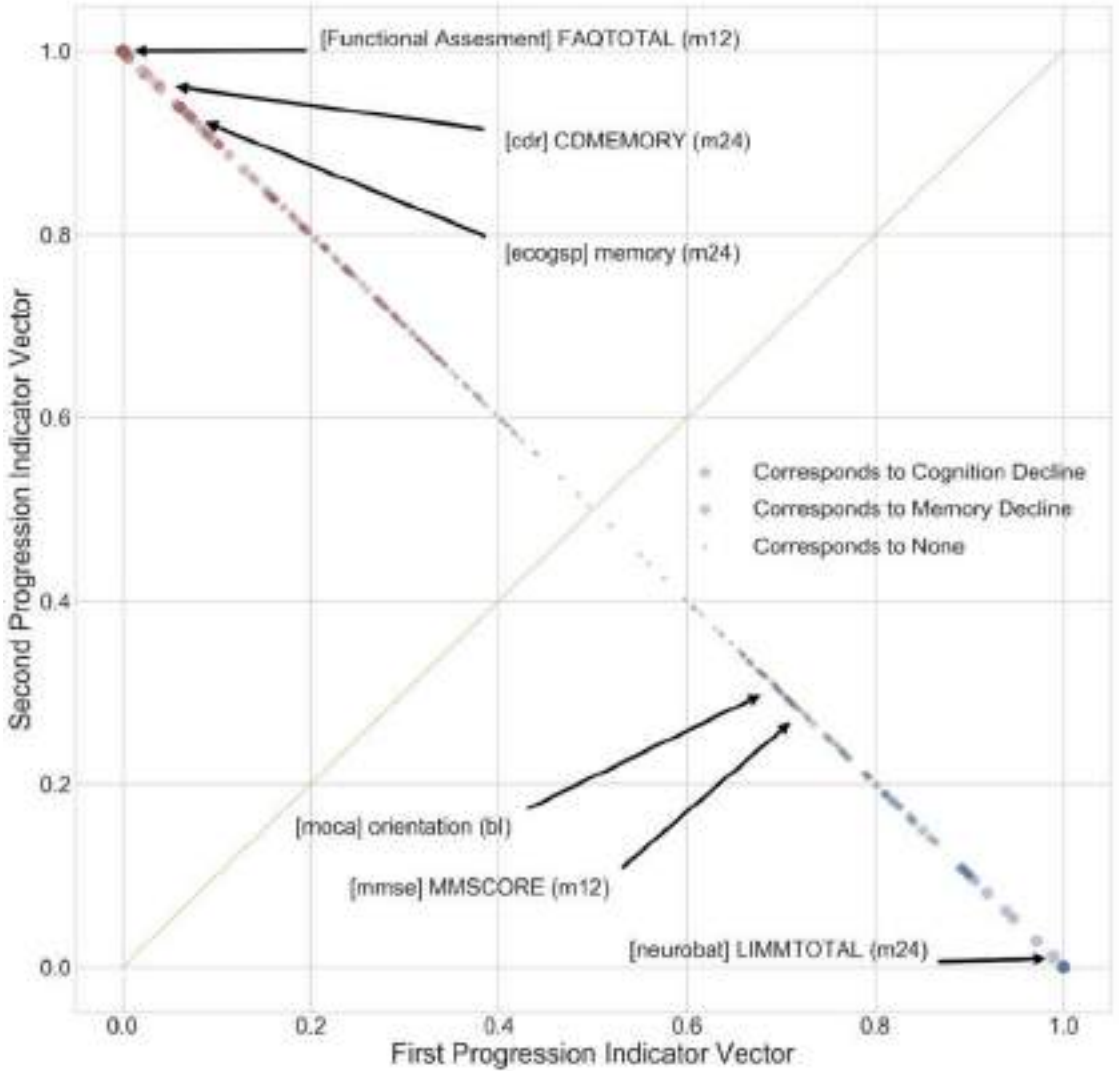

Figure 6.2: The plot of features in two dimensions using progression indicator vectors.
Features in red correspond to memory decline, and features in blue correspond to global cognitive decline. Yellow line with a slope of 1, which separates the features into two categories, is drawn for reference. Features that occur below this separating line were associated with cognitive decline (x-axis) in the AD projection space, and features that lie above the 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.

6.3.3 Details on AD Progression Space

In this section, we discuss the following additional details on the AD progression space:

1) reversion of AD captured in the constructed progression space (Section 6.3.3.1),

2) correlationbetweentheAPOEε4geneticvariantsandparticipant'sprogressionstate(6.3.3.2),

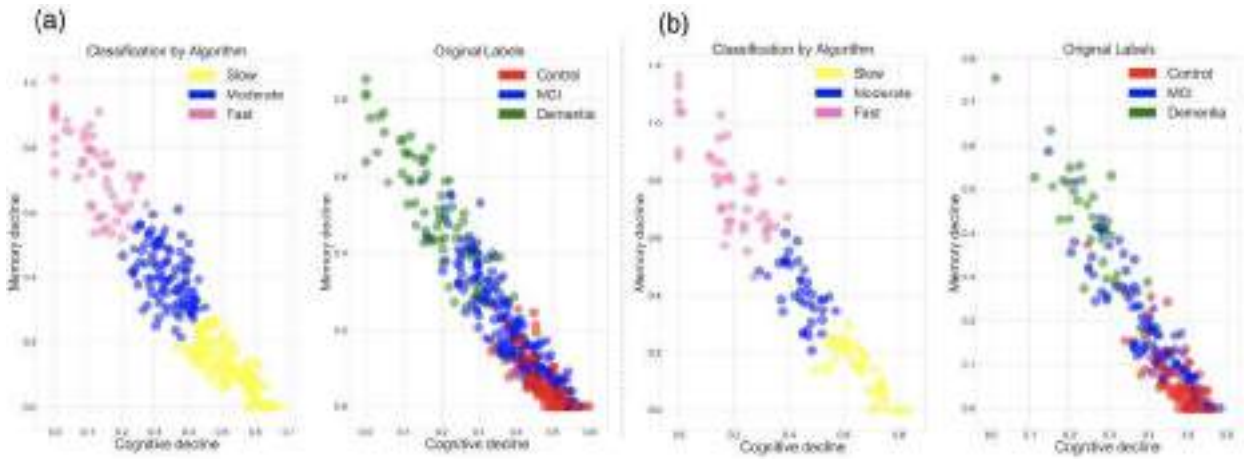

Figure 6.3: Views of AD progression space (a) Comparison of 24th month machine learning
based prediction and original labels. A total of 582 cases are projected in the AD progression space at the 24th month. Left: Machine learning based classification. Slow, moderate and fast progression rates are represented in yellow, blue and pink, respectively. Right: Colored with original labels. Controls, MCI, and dementia patients are represented in red, blue and green, respectively. (b) Comparison of 48th month machine learning based prediction and original labels. A total of 253 cases are projected in the AD progression space at the 48th month. Left: Machine learning based classification. Slow, moderate and fast progression rates are represented in yellow, blue and pink, respectively. Right: Colored with original labels. Controls, MCI, and dementia patients are represented in red, blue and green, respectively.

Table 6.1: Percent share of diagnostic group and 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.

		Subtype at 24 months from baseline		
Outcome	Controls	Slow	Moderate	Fast
MCI Dementia	-	100%	69.39 %	9.43%
APOEε4 count = 0	-	0%	10.82 %	90.57%
APOEε4 count = 1	72.22%	61.40%	30.61 %	28.30%
APOEε4 count = 2	25.93%	31.58%	15.31 %	50.94%
	1.85%	7.02%	43.88 %	20.75%

3) effects of aging on AD progression in controls (6.3.3.3),

4) correlation of memory decline in AD patients with their educational and occupational attainments (6.3.3.4),

5) distribution of projected dimensions (memory decline and cognitive decline) for each AD

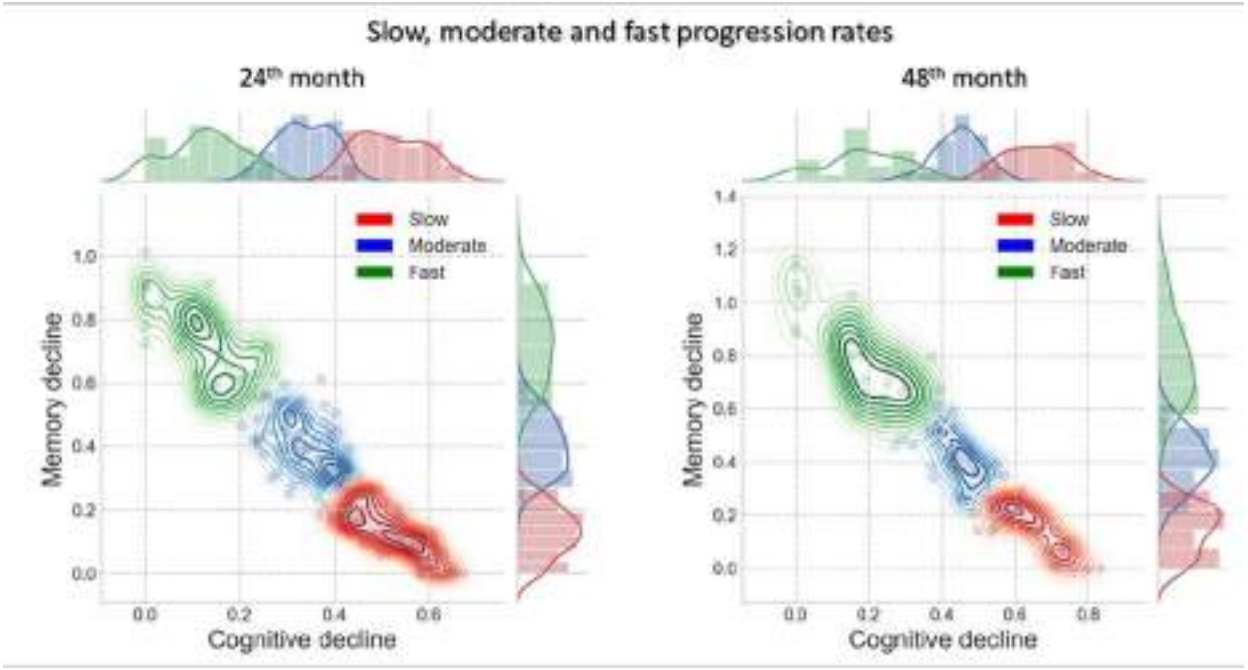

Figure 6.4: Three different progression rates are identified in MCI and dementia patients. Left: at the 24th month. The slow, moderate and fast progression rate zones are represented in red, blue and green, respectively. Right: at the 48th month. The slow, moderate and fast progression rate zones are represented in red, blue and green, respectively.

subtype (6.3.3.5), and

6) correlation between certain selective features and AD progression rates (6.3.3.8).

6.3.3.1 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 6.6 plots these reversion cases at the 24th and 48th 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.

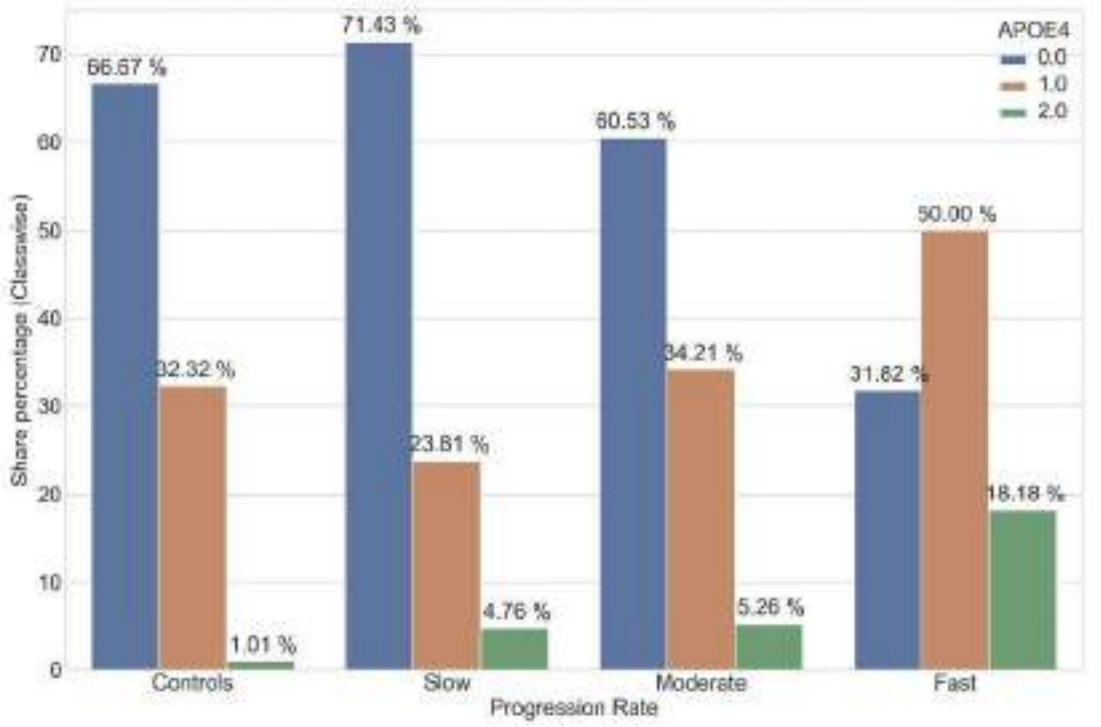

Figure6.5: Percent share of APOEε4 variants for different subtypes after 48 months from baseline. The share of 0 occurrences of APOEε4 variants is decreasing with increase in progression rate, whereas the share of other two 1 and 2 occurrences is increasing.

6.3.3.2 ADprogressionstatesandAPOEε4allelecounts

To understand the underlying biological patterns among patients in the progression space, we plotted the distribution of the APOEε4 alleles. Figure 6.7 projects the observations with 0 and 2 counts of APOEε4 variants on the AD progression space at 24th and 48th month. It is evident from these figures that observations with a 0 count for the APOEε4 allele are concentrated towards the slow progression rate zone, whereas observations with 2 counts of APOEε4 allele are concentrated towards the moderate and fast progression rate zones. This observation further validates the existing literature [214] and confirms a significant correlation between APOEε4 with decreased cognitive performance.

6.3.3.3 ADprogressionincontrolsandaging

In Figure 6.8, progression can also be seen in control observations at 24th and 48th 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 fast

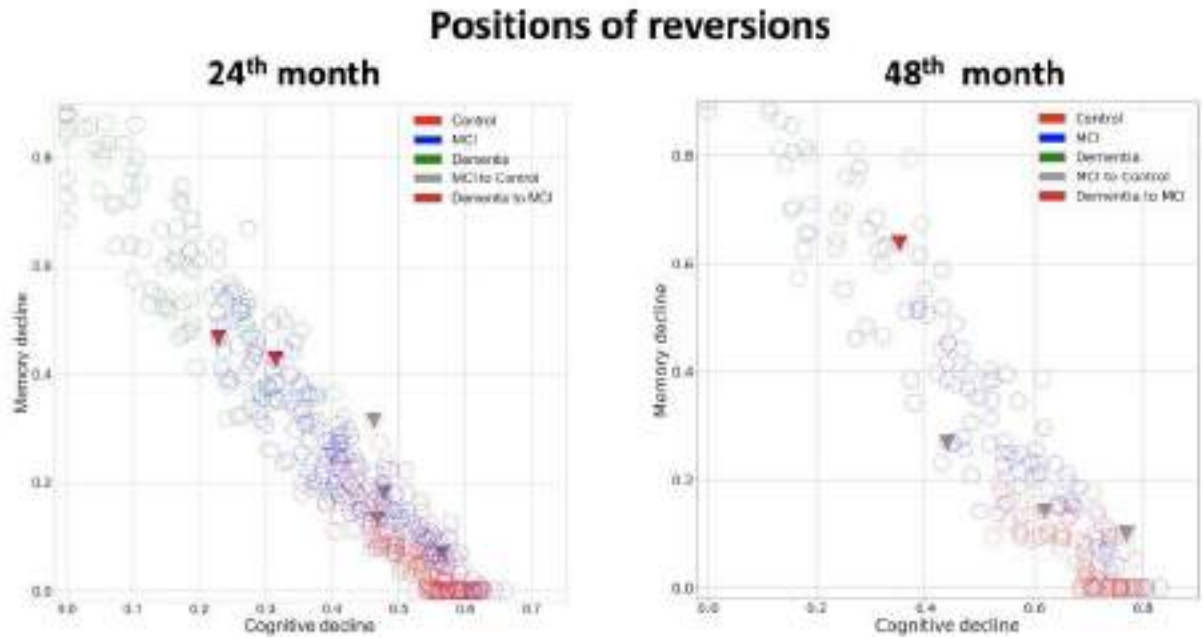

Figure 6.6: Label Reversions (only MCI to control and dementia to MCI). Left: The positions of the reversions at 24th month relative to all other observations. Right: The positions of the reversions at 48th month relative to all other observations.

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 in Figure 6.8 that at the 24th month, the mean age of the cluster which is relatively close to the moderate progression rate zone is 75.15 years (95% confidence interval (CI): 73.42-76.90), and the mean age of the cluster away from this zone is 71.98 years (95% CI: 70.92-73.05). Similarly, at the 48th month, the mean ages of the two clusters relatively close and away from the moderate progression zone are 71.85 (95% CI:70.34-73.37) and 73.39 (95% CI: 70.91-75.87), respectively.

6.3.3.4 MemorydeclineinADpatientsandeducationalattainment 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 6.9 (Left), the cubic curve fits the data in a linear fashion in the slow progression region. However, it deviates slightly from this linear behavior in fast and moderate progression regions. 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 6.9 (Right). The magnitude of slope for the linear curve is 1.26 and 1.19 for highly educated and less educated

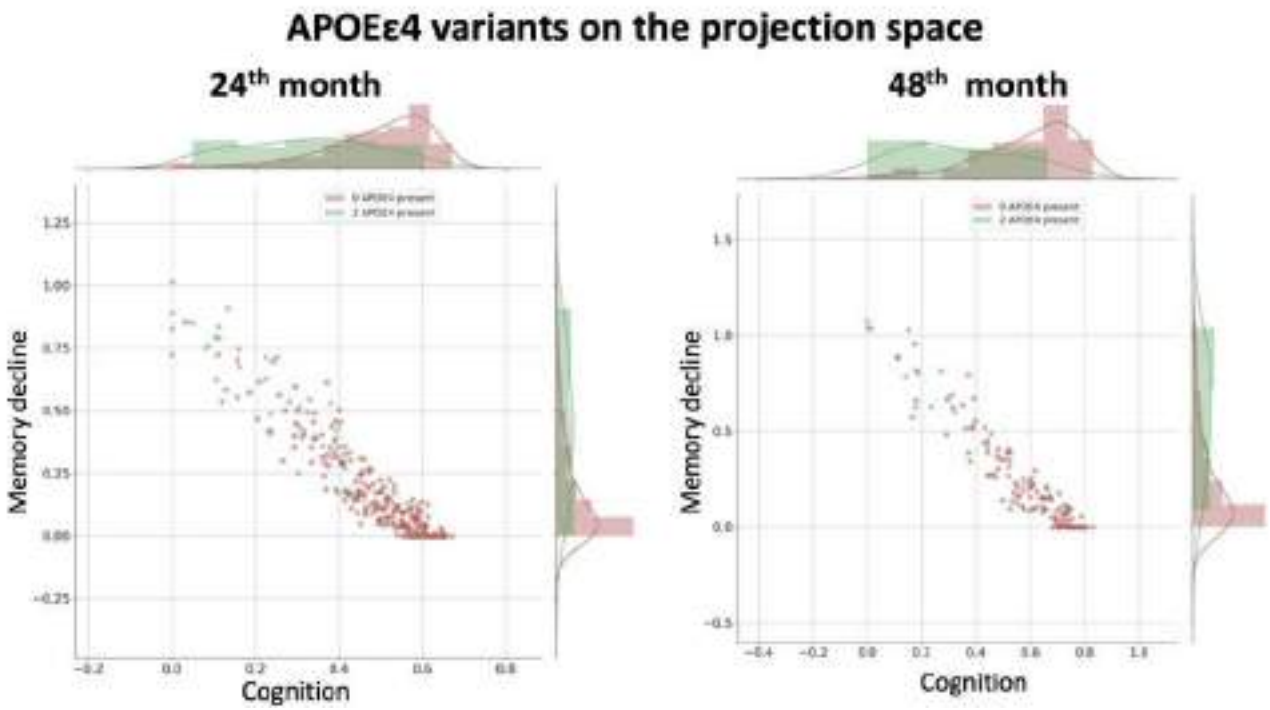

Figure 6.7: Projection of the number of APOEε4 variants on the projection space. x-axis and y-axis have visualizations of the distribution of APOEε4 alleles in those directions. Left: at 24th month. Right: at 48th month.

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 [215]. 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 [215].

In the above analysis, we have seen correlation between age, education and APOEε4 variant with progression rate. But just these three variables are insufficient to predict progression rate. While accuracy and AUC for prediction using longitudinal clinical data is above 80%, these values are far less if only age, education, and APOEε4 are used (see Figure 6.10).

After using only age, education and APOEε4, data prediction accuracy is 47.54 (95% CI, 43.84-51.24) and AUC for slow, moderate, fast is 0.486 (95% CI, 0.443-0.530), 0.627 (95% CI, 0.535-0.719) and 0.679 (95% CI, 0.625-0.734), respectively (Figure 6.11).

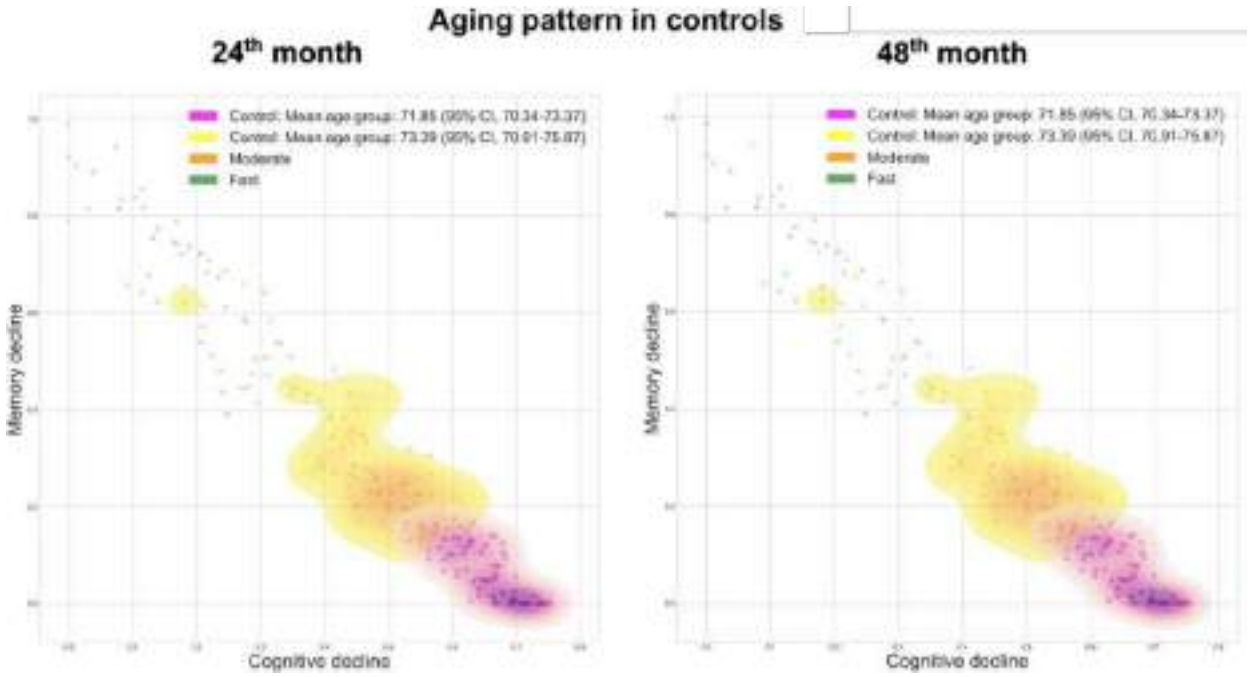

Figure 6.8: Aging pattern in controls. Mean age of cases in the two clusters of controls. Clusters represent an aging pattern in controls. The cluster near a moderate progression rate zone has a higher mean age than one which is away from it. Left: at the 24th month. Right: at the 48th month.

6.3.3.5 Distributionofmemoryandcognitivedecline Figure 6.12 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 higher cognitive decline whereas a high value on the y-axis indicates higher memory decline. Fast progression rate has the highest memory and cognitive decline, which goes on reducing with a reduction in progression rate.

6.3.3.6 Supervisedsubtypeprediction

The progression rate of participants after the 24th and 48th months from the baseline were predicted using the random forest classifier. It gave the best five-fold cross-validation accuracy and AUC curve results for all the cases. A comparison of different machine learning algorithms for the prediction of progression after 24 and 48 months from the baseline is given in Figures C.3 and C.4 in the Appendix, respectively. For the 24th month using baseline and 12 months of ob-

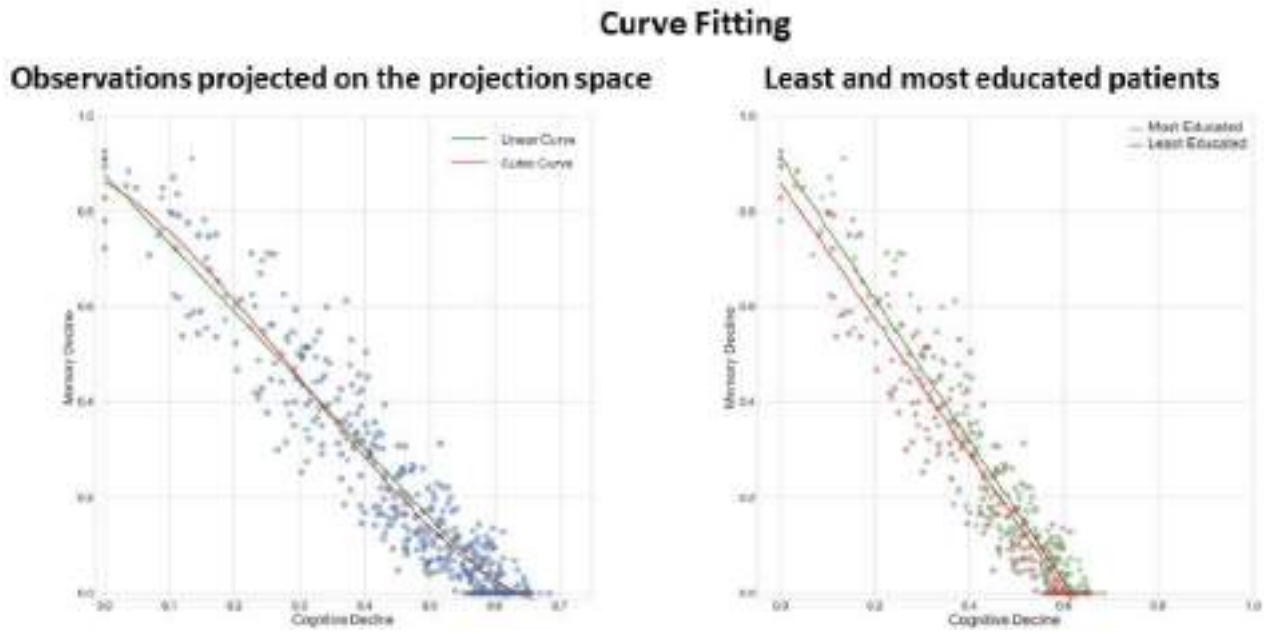

Figure 6.9: Memory decline in AD patients and educational attainments. Left: Linear and cubic curve-fitting on 582 observations projected on the progression space. The magnitude of the slope of the linear curve is 1.19 indicating a rapid memory decline as compared to cognitive. Right: Linear curve fitting for 200 most educated and 200 least educated patients.

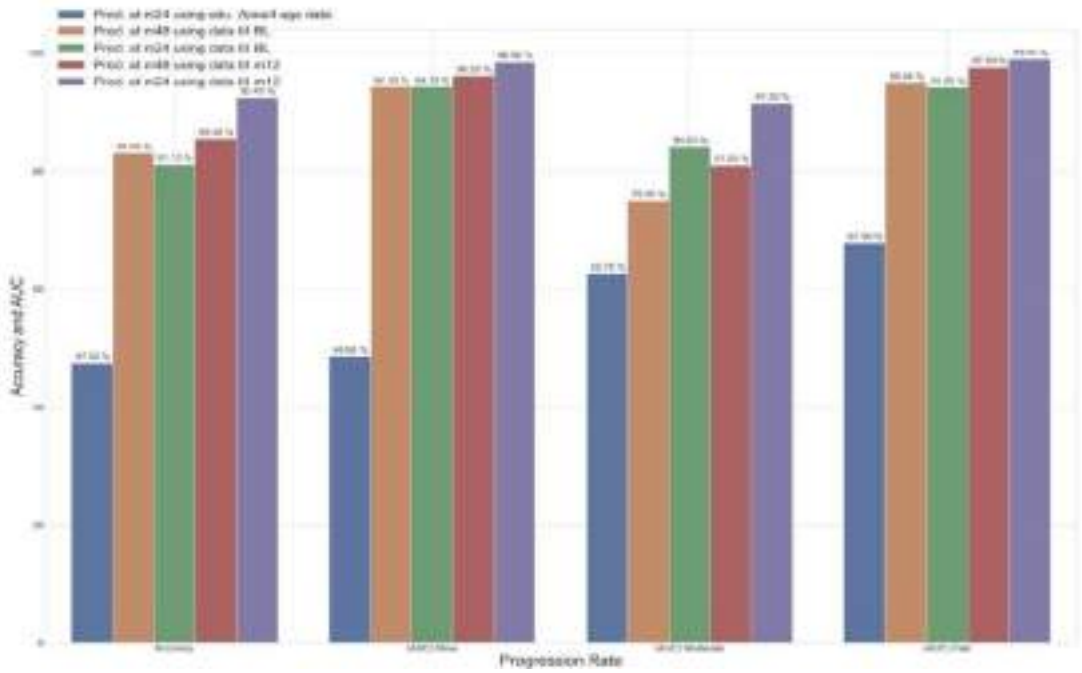

Figure 6.10: Comparing accuracy and AUC for prediction using only age, education, and APOEε4 as predictors with baseline and 1 year longitudinal clinical data.

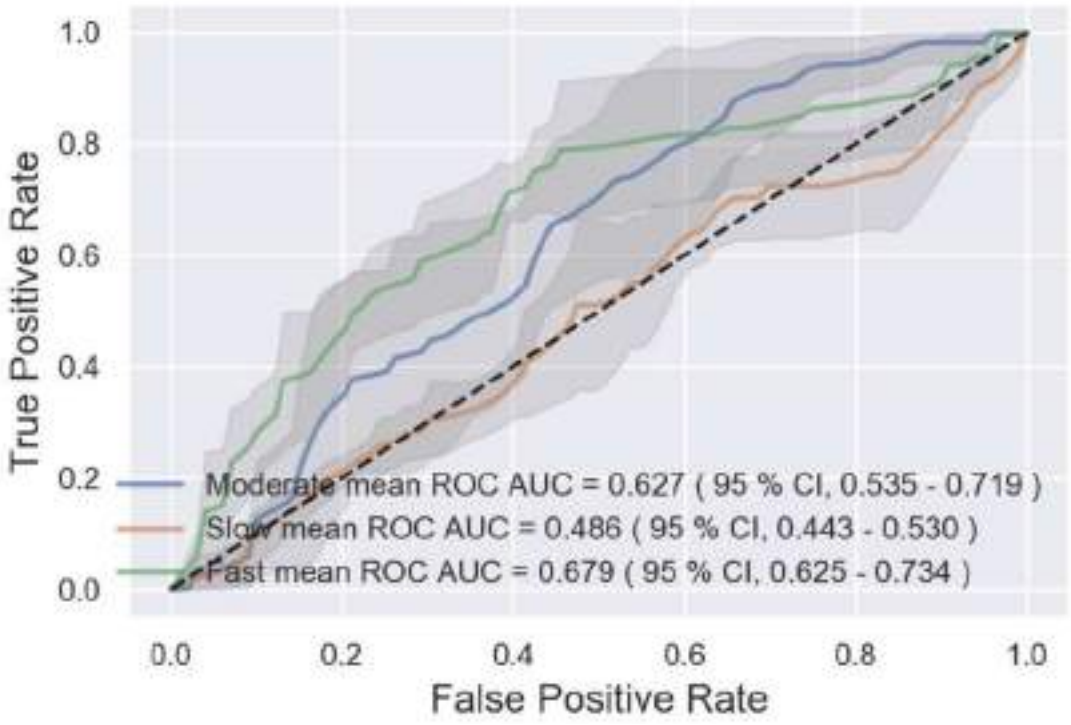

Figure 6.11: Area under ROC curve for predictions using age, education and APOEε4 data.

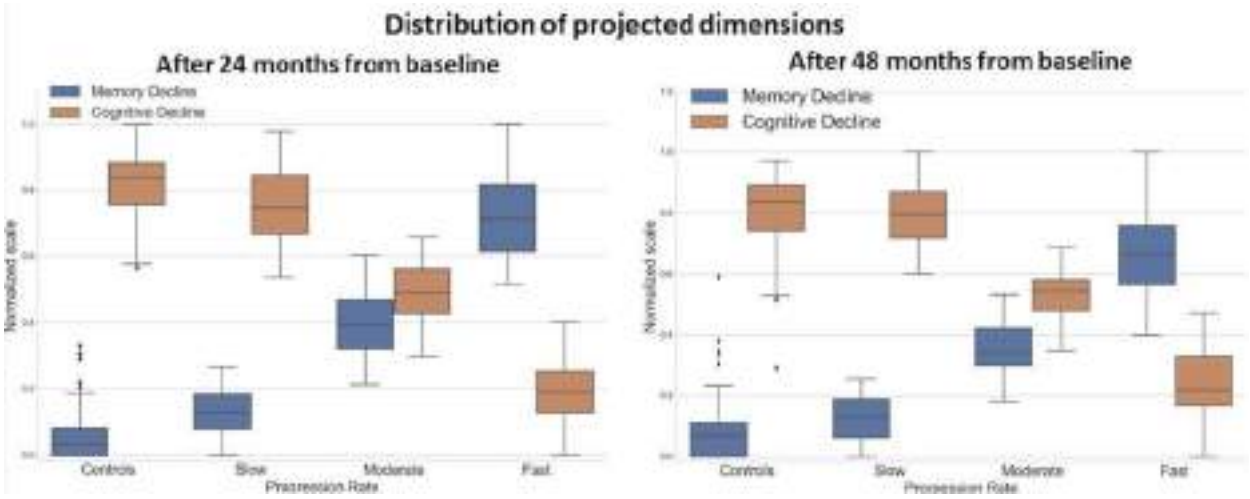

Figure 6.12: Distribution of projected dimensions (cognitive and memory decline) for each AD progression subtype. A lower numeric value on the cognition axis indicates fast cognitive decline whereas a higher numeric value on the memory axis indicates fast memory decline. A similar relationship can be observed in the figures. Left: after 24 months from baseline. Right: after 48 months from baseline.

servations, AUC of 0.98 (95% CI, 0.976-0.994), 0.915 (95% CI, 0.885-0.946) and 0.990 (95% CI, 0.984-0.997) for slow, moderate and fast progression rates were observed, respectively. Pre-

diction of progression at 48th month using baseline and 12 months of observations yields AUC of 0.962 (95% CI, 0.931-0.994), 0.810 (95% CI, 0.750-0.870) and 0.976 (95% CI, 0.961-0.991) for slow, moderate and fast progression rates, respectively. Note that we used the "one vs rest" paradigm to compute the above reported multiclass AUCs, i.e., we iteratively considered each class as our "positive" class and compared it against all the other classes at the same time ("negative" class). This is similar to reducing our multiclass classification problem into three binary classification ones. In our implementation, the accuracy for the prediction of progression at the 24th and 48th month is 92.45% (95% CI, 91.35%-93.55%) and 85.48% (95% CI, 79.81%-91.16%), respectively (Table 6.2). Figure 6.13 depicts the ROC curves for three separate classes

Table 6.2: The performance of the optimized machine learning model in predicting subtypes observed at 24th and 48th month from baseline using baseline and one year of data.

Progression month	Accuracy (95% CI)	AUC (95% CI)		
		Slow	Moderate	Fast
Predicting subtype observed at 24th month using baseline data	0.811 ± 0.063 (0.748-0.874)	0.943 ± 0.030 (0.913-0.973)	0.842 ± 0.051 (0.791-0.893)	0.942 ± 0.033 (0.909-0.975)
Predicting subtype observed at 48th month using baseline data	0.831 ± 0.079 (0.752-0.91)	0.943 ± 0.025 (0.918-0.968)	0.750 ± 0.063 (0.687-0.813)	0.950 ± 0.028 (0.922-0.978)
Predicting subtype observed at 24th month using one year data	0.924 ± 0.011 (0.913-0.935)	0.985 ± 0.009 (0.976-0.994)	0.915 ± 0.030 (0.885-0.945)	0.990 ± 0.006 (0.984-0.996)
Predicting subtype observed at 48th month using one year data	0.855 ± 0.056 (0.799-0.911)	0.962 ± 0.031 (0.931-0.993)	0.810 ± 0.06 (0.750-0.87)	0.976 ± 0.015 (0.961-0.991)

(slow, moderate and fast progression rates) for the 24th and 48th month.

LSTM model was also used for the prediction of AD subtypes with slow, moderate and fast progression rates as well as controls. The accuracy of prediction of projection rates using LSTM is 82.58% (95% CI: 81.00 - 84.16) and 87.40% (95% CI: 86.02 - 88.78) for the 48th and 24th 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. Further, we also built a trajectory prediction model using LSTM networks. See Section C.2.3 in the Appendix for details.

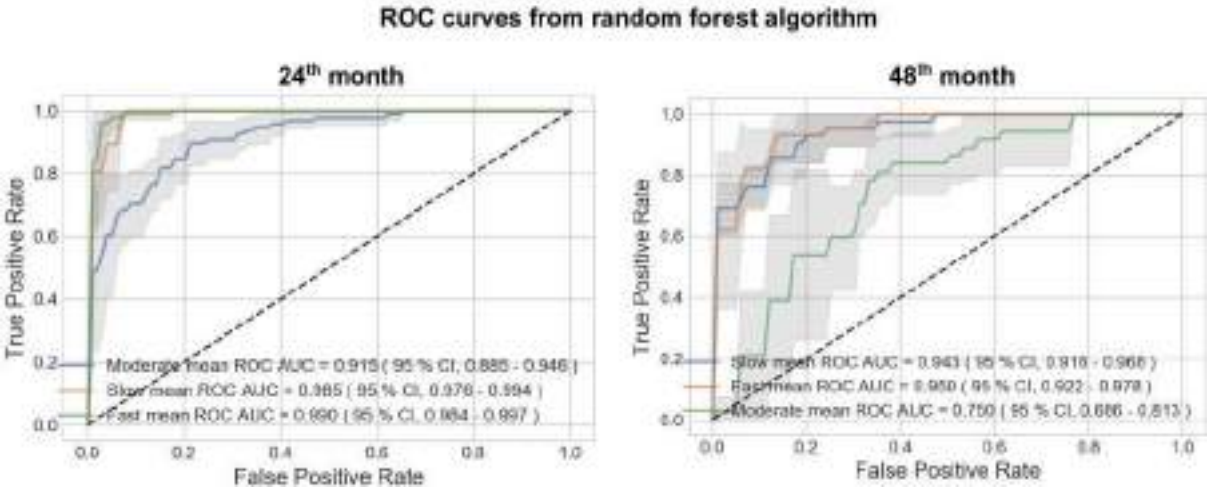

Figure 6.13: ROC of AD patient's progression rate after the 24th and 48th months based on the one year and baseline data (resp.), including the area under the ROC for three AD progression subtypes (slow, moderate and fast progression rates). Left: The predictions for thediseasestageatthe24thmonthweremadeusingarandomforestalgorithm. Right: The predictions for the disease stage at the 48th month were made using a random forest algorithm.

6.3.3.7 Supervisedmodelinterpretation We used the Shapley additive explanations (SHAP) approach to evaluate each clinical feature's influence in ensemble learning. This approach is used in game theory and assigns an importance (i.e., SHAP) value to each feature to determine a player's contribution to success. SHAP enhances understanding by creating accurate explanations for each observation in a dataset. Figure 6.14 highlights the distribution of the top features that were used to predict the identified AD subtypes. The interactive website was developed as an open-access, to provide a simple-to-use tool that clinicians can access.

6.3.3.8 SelectivefeaturesandADprogressionrates

Figure 6.15 (left) depicts the distribution of the MMSE score after 6 and 12 months for each AD subtype at the 24th 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 48th month (given in Figure C.6 in the Appendix). Further, there is an increase in functional assessment questionnaire (FAQ) total score with increasing progression rate for the 24th and 48th month AD subtypes as shown in Figures 6.15 (right) and C.7, respectively.

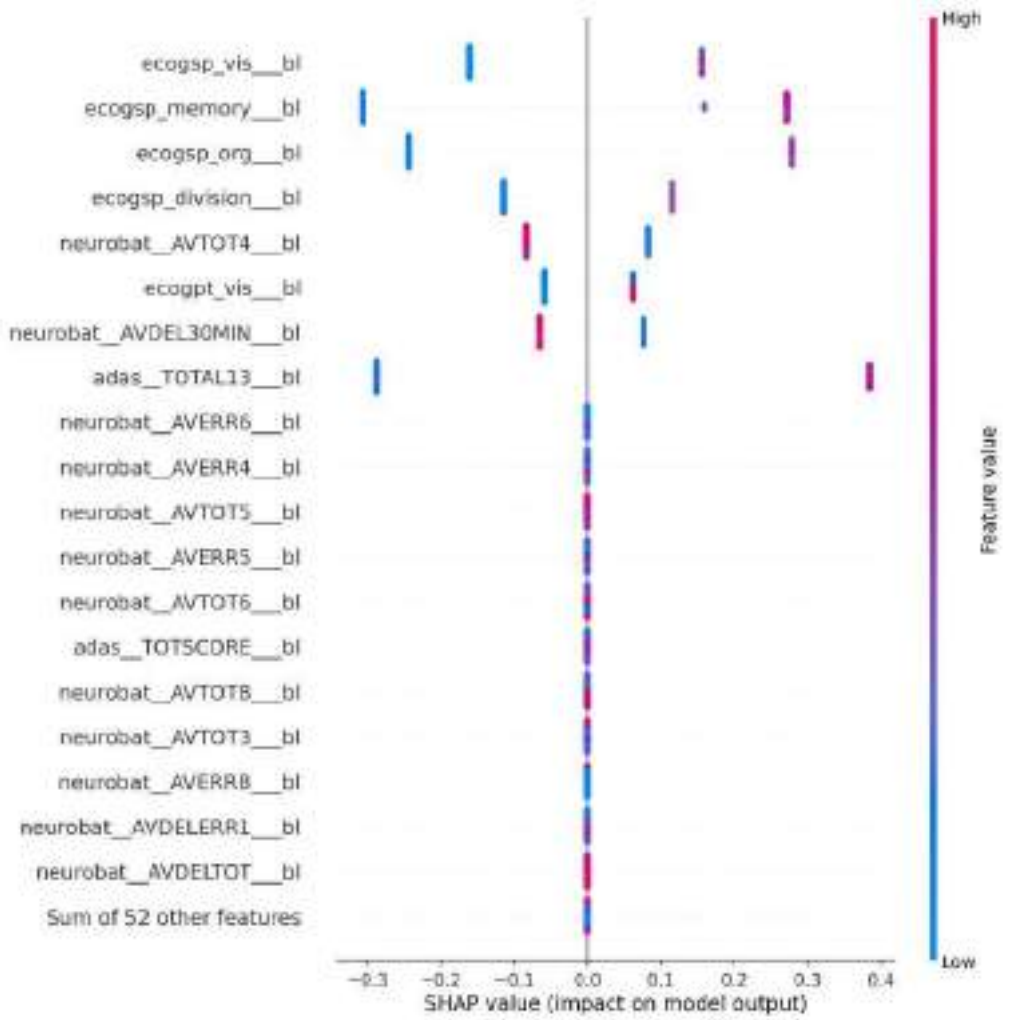

Figure 6.14: Distribution of the top features that had the most substantial effect on the predictive value of the classification model over all subtype classes. Each point represents a patient and the amount of effect on model output for each feature depends on its SHAP value. For example, the effect of the ADAS score (adas_TOTAL13_bl) on model output is large (severe AD subtype) when the patient has high values for the ADAS feature (in red) as compared to its low values (in blue).

6.4 Discussion

This work clusters participants in distinct progression stages of AD and discusses an approach to predict the future rate of progression after the 24th and 48th months from baseline using longitudi- nal clinical data. Predicting disease progression serves as a paramount challenge for the develop- ment of disease-modifying early-intervention therapies of complex neurodegenerative diseases. This study is a step towards designing sophisticated machine-learning paradigms that facilitate accurate prediction of AD progression. Such predictions would lead to better patient-specific

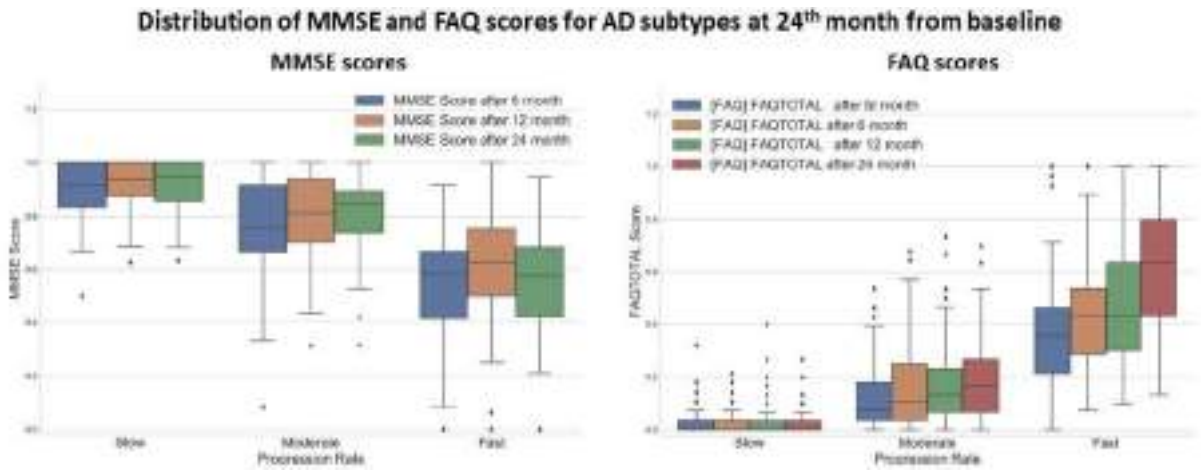

Figure 6.15: Left: Distribution of the normalized MMSE score between 0 and 1 for each AD subtype (subtypes at 24th month). MMSE score decreases with an increase in progression rate. Right: Distribution of FAQ total score for each AD subtype (subtypes at 24th month). FAQ total score increases with an increase in progression rate. Abbreviations: MMSE: mini-mental state examination; FAQ: functional activities questionnaire; and ADAS: advanced driver assistance systems.

attention by recognizing at an early stage the patients with a swift rate of progression. The proposed disease progression and trajectory prediction algorithms can be particularly helpful in stratifying patient cohorts recruited for clinical trials. Further, no disease-modifying treatments in AD exist, hence, the capability to anticipate the trajectory of impending AD progression at the early and prodromal stages of the disease is an advancement towards uncovering novel treatments for AD modification. The study provides insights into the progression of AD-related symptoms and signs.

In this work, we discussed the share of different APOEε4 alleles for each progression rate and its correlation with cognitive performance. Future work should involve examining additional genetic factors that also have been closely related to the progression of AD for studying their interactions, such as polygenic risk scores [216, 217]. As stated earlier, AD risk is associated with APOEε4 status [213]. 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 [218]. 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 [217].

The present analysis can be continued in various directions. This study involved 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 cross-validation to gauge the performance of our models, validating the study on separate AD datasets is required [219, 220]. Additionally, the discussed analysis only focused on predicting progression space in AD. The proposed framework can be further adapted to study additional forms of dementia, such as frontotemporal degenerations, Lewy body demen- tia, and multi-infarct dementia. Finally, we hope to apply and expand this project in the future, leveraging larger datasets with more diverse participants.

6.5 Summary

AD is a common, age-related, neurodegenerative disease that impairs a person's ability to perform basic activities of daily living. Diagnosing AD can be challenging, especially in the early stages. Many patients remain undiagnosed, partly due to the complex heterogeneity in disease progression. This diagnostic challenge highlights a need for early prediction of the disease course to assist its treatment and tailor management to the disease progression rate. Recent developments in machine learning techniques provide the potential to predict disease progression and trajectory of AD and classify the disease into distinct subtypes.

The work shown here clusters participants in distinct and multifaceted progression subgroups of AD and discusses an approach to predict the progression rate for each subgroup from the baseline diagnosis. We observed that the myriad of clinically reported symptoms summarized in the proposed AD progression space correspond directly with memory and cognitive measures, routinely used to monitor disease onset and progression. Our analysis demonstrated accurate prediction of disease progression after four years from the first 12 months of post-diagnosis clinical data using five-fold cross-validation (area under the curve of 0.96 [95% CI, 0.948-0.977], 0.81 [95% CI, 0.783-0.837] and 0.98 [95% CI, 0.969-0.983] for slow, moderate and fast progression rate patients, respectively). Further, we explored the long short-term memory neural networks to predict an individual patient's progression trajectory.

The machine learning techniques presented in this study may help providers identify distinct disease subtypes with different progression rates and trajectories in the early stages of the disease, allowing for more efficient and personalized healthcare delivery. With additional information about the progression rate of AD at hand, providers may further individualize treatment plans. The predictive tests discussed in this study allow for early AD diagnosis and facilitate the characterization of distinct AD subtypes relating to disease progression. These findings are

CHAPTER 7

EXAMINING NEURAL AND PHYSIOLOGICAL RESPONSES TO BALANCE-DEMANDING TARGET-REACHING LEANING TASKS

In this chapter, we review the work in Using Virtual Reality to Examine the Neural and Physiological Anxiety-Related Responses to Balance-Demanding Target-Reaching Leaning Tasks. This work previously appeared as [79]. The code for this work is publicly available

7.1 Introduction

Stress, being an autonomic nervous system reaction to threats and challenges [221], is known to interact with task performance. Excessive stress has been demonstrated to negatively affect human performance in demanding tasks both physically [222] and cognitively [223], and a bad task performance in turn induces further stress [224], [225]. Anxiety, being an emotion characterized by an unpleasant state of inner turmoil [226], has been shown to interact with postural instabil- ity [227], [228]. Further anxiety, in general, can be classified into trait anxiety and state anxiety. Trait anxiety is a relatively stable personality characteristic that does not change much from day to day, while state anxiety is transitory and situationally bound [229], [230]. Patient-rated dis- ability has been shown to exaggerate fear of falling (FoF), a form of trait anxiety disorder [231]. At the same time, FoF associated with upright balance can cause stiffness [50], elevating the possibility of falling and consequently raising the FoF. However, fewer studies have been done in the past on how state anxiety and motor task performance interact.

In the proposed work, a novel virtual reality (VR) based test-bed was examined to study the neural and physiological anxiety-related responses in a balance-demanding target-reaching whole body leaning task. To analyze the subjects' neural and cardiac responses in real time during the experiments, electroencephalography (EEG) and electrocardiography (EKG) data collection systems were established. The brain-computer interface (BCI) system, integrating VR and online anxiety level feedback, provides the technology to study anxiety and motor task performance in clinical or industrial settings, and training opportunities to improve postural control. The proposed framework is an advancement progressing towards an intuitive human-robot

or BCI system that can detect anxiety in human users during demanding motor tasks, and in anxiety-inducing industrial work environments. Moreover, in this study, machine learning (ML) is applied to validate the effectiveness of using frontal alpha band in predicting state anxiety detected by heart rate variability (HRV) as the ground truth. HRV is a suited biomarker used by researchers for identifying anxiety, psychological symptoms and stress-related diseases [232]. The advancement of these prediction models through ML may help clinicians with decision making and developing personalized clinical care and counseling.

7.1.1 Objective

We aim to validate the use of frontal alpha band and frontal alpha asymmetry (FAA) from EEG signals in measuring anxiety levels. To accomplish the same, we refer to HRV from the EKG signals and Patient-Reported Outcomes Measurement Information System (PROMIS) trait anxiety questionnaire [233] assessing anxiety encountered in daily life, self-stated out by the subjects, as the ground truths. We also want to understand the dynamics of anxiety in a challenging motor task by studying the accuracy and speed of the subjects real-time task performance. Figure 7.1 indicates a conceptual map for our proposed study.

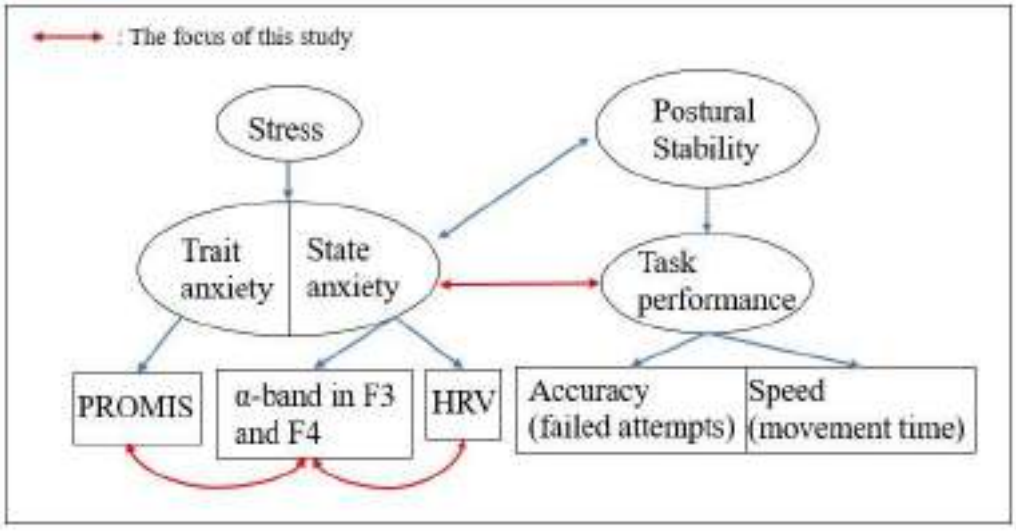

Figure 7.1: Conceptual map for our analysis

7.1.2 Related Work

Although reduced HRV was first brought up as a physiological indicator of anxiety [234], no universal standard for anxiety evaluation has yet been recognized [235]. It has been shown that HRV is lower in populations with anxiety disorders [236] and when state-anxiety is induced in stress- ful tasks [237]. Another potential bio-marker widely studied for detecting anxiety is FAA [238]. To investigate anxiety and neurofeedback, researchers have been exploring the integration of VR environments and EEG. Simulated high postural threat conditions were induced in an immer- sive VR environment so as to minimize fall-risks in elderly adults [61]. In an effort to train subjects to control anxiety levels in balance-demanding walking conditions, a self-monitoring and responsive VR-based test-bed was studied by analyzing the real-time neurological feedback (recorded via EEG) to acrophobic responses (induced via visual stimuli) in VR [63]. An exper- imental setup configured to inspect neural responses while subjects undergo randomized height changes and perturbations in a quiet standing virtual environment was proposed in [78]. Several other VR and BCI-based systems have been also studied for rehabilitation [70], [71], [69] by researchers. To further analyze features in EEG signals, ML has been explored. EEG signals were used to detect human stress through a K-Nearest Neighbor (KNN) classifier [239] and to classify subjects into groups of high and low hypnotic susceptibility [240]. Recent developments in ML techniques also provide a huge potential to classify and predict the onset and progression of anxiety [241]. To the best of our knowledge, our study is the first that applies ML to validate the use of frontal alpha band in detecting state anxiety, and studies the interaction between state anxiety and motor task performance, while adapting a VR-based experimental test bed.

The remainder of the chapter is organized as follows. In Section 7.2 and 7.2.3, we present the methodology for data collection, EEG and EKG data processing, and the EEG signal clas- sification approach using ML algorithms. In Section 7.3, we discuss major results and examine some applications of our VR-based test-bed. Finally in Section 7.4, we highlight the concluding remarks and future directions for our work.

7.2 Methods

The protocol for the study was approved under the Institutional Review Board number 17010.

7.2.1 Experimental Setup and Data Collection

To assess the effects of anxiety on a challenging motor task, we set up an EEG and EKG data collection system (as represented in Figure 7.2) consisting of the following:

1) a HTC Vive VR headset

2) an AMTI AccuSway force plate

3) a Brainvision 64 Channel ActiCHamp EEG system

4) a Trigno EKG biofeedback sensor

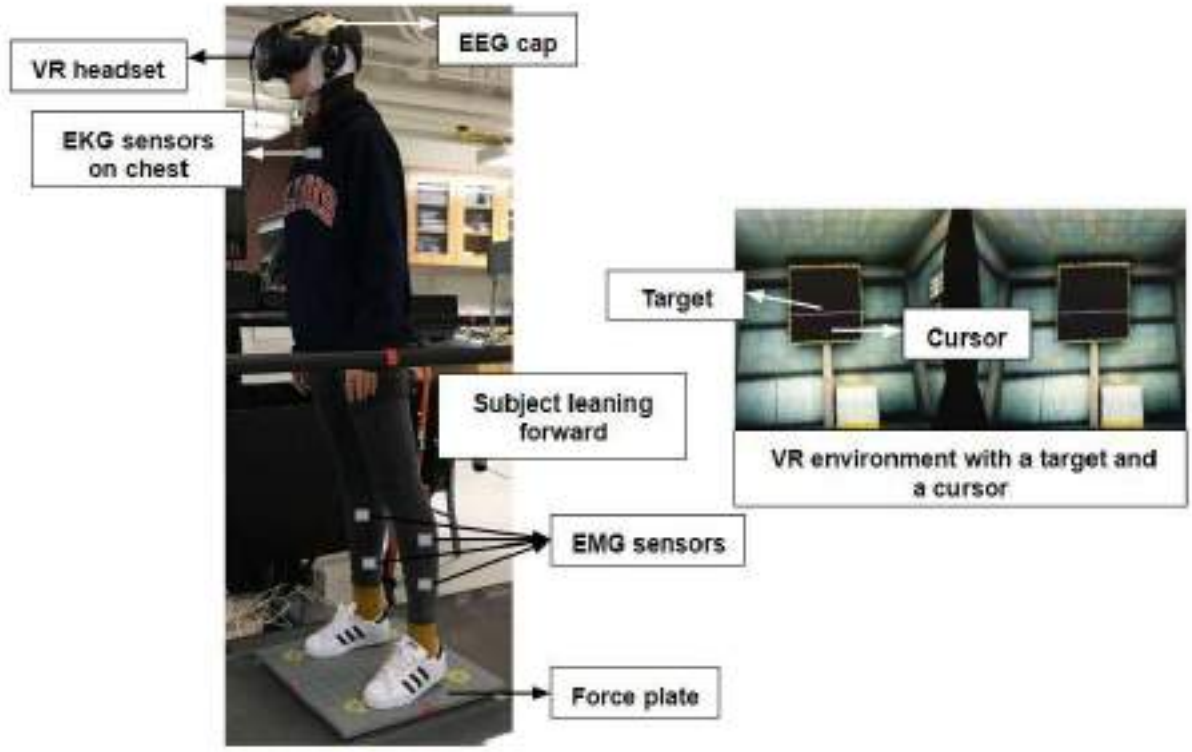

Figure 7.2: VR Leaning experiment: Data collection setup and virtual reality environment. All subjects stood on an AMTI portable force plate throughout all the trials. In the virtual environment, subjects saw a platform that aligned with the physical force plate they were standing on. With a virtual cursor that moves up and down by leaning forward and backward respectively, all subjects were asked to strike a target in virtual reality until it disappeared (right). During half of the trials, an ocular feedback of significant height change between the platform and the floor was demonstrated while a levelled surface was demonstrated during the other half.

All subjects experienced a virtual world where they stood on a platform. During baseline recordings, subjects were instructed to stand quietly on the force plate for 5 seconds. They were

then asked to lean forward as far as they could by dorsiflexing the ankle joints to capture the maximum forward lean condition. The average center of pressure on the y-axis was recorded by the force plate in both the baseline and maximum forward lean conditions. The difference between the two average center of pressures is defined as the maximal forward lean distance (MFLD). Subjects were provided a virtual cursor that they could move up or down by leaning forward or backward, respectively, and a target in VR. The distance between the start position of the cursor and the target indicated 84% of the MFLD in half of the experimental trials, and 56% in the other half. The MFLD is re-measured before each trial. During the trials with 56% MFLD, visual feedback indicating a significant change in height between the platform and the floor was presented. This was considered as a high fall-risk condition for our study. Visual feedback of a levelled surface was provided for the trials with 84% MFLD. During the experiment, subjects were asked to move the cursor to the target by leaning forward and staying on it until it disappeared (in 200 milliseconds (ms)). The target area spanned over a width of 20% of the target distance. If the cursor dwelt in the target area for 200 ms, the target disappeared, indicating a successful acquisition. If it fell out of the target area within 200 ms, the attempt failed. For a failed attempt, subjects were instructed to retry and move the cursor back into the target area. A total of 240 targets were carried out in 30-60 target increments (i.e. 8 and 4 experimental trials respectively) for each subject. Figures 7.3 and 7.4 depict the details of virtual environment and target acquisition task, respectively, used in this study.

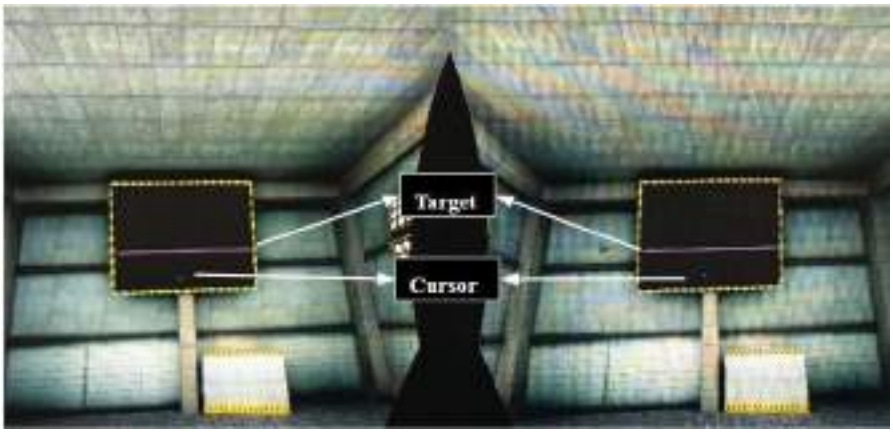

Figure 7.3: VR environment for the target acquisition task. Visual feedback indicating a significant change in height between the platform and floor was presented during the trials with 56% MFLD and that of a levelled surface was provided for trials with 84% MFLD. One target is shown, but 30 targets are presented one at a time at each trial.

During each trial, high-density EEG data from 64 channels of the brain and EKG signals from the heart were recorded. To analyze the real-time anxiety states, EEG and EKG signal processing pipelines were implemented using Python 3.6 with the support of several open source libraries

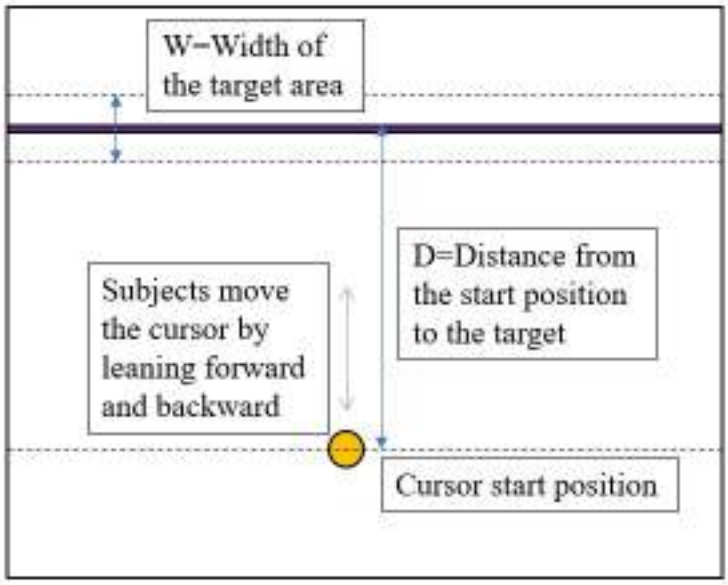

Figure 7.4: Target acquisition task. One target is shown, but 30 targets are presented one at a time at each trial. In half of the trials, the distance is 84% of the maximum lean distance. In the other half of the trials, the distance is 56% of the maximum lean distance. The width of the target area is always 20% of the target distance. The maximum lean distance is re-measured before each trial.

(MNE, scikit-learn, pandas etc.) and MATLAB respectively.

7.2.2 Subject Demographics

Data from 10 healthy young adult (HYA) volunteers from the local community (with 5 (50%) females) was used for this study. Subjects included for the study were right-side dominant with a normal or corrected to normal vision. Further, subjects who self-reported neurological, vestibular or musculoskeletal conditions were excluded from the data collection process. A summary of descriptive statistics on the demographic information of subjects is given in Table 7.1.

Table 7.1: Descriptive Statistics on demographic information

Cohort	Age(years)	Weight(kg)	Height(cm)	BMI(kg/m2)
HYA	21.7 ±1.9	70.4±10.9	171.5±7.6	23.8±2.3

7.2.3 Data Analysis

A brain computer interface approach was used while investigating the EEG and EKG signals to analyze instantaneous anxiety.

7.2.3.1 EEGprocessing

During the course of the entire experiment, EEG signals at a sampling rate of 1000 Hz were recorded from 64 electrodes on the subjects' scalp. See Figure 7.5. Six external electrodes were

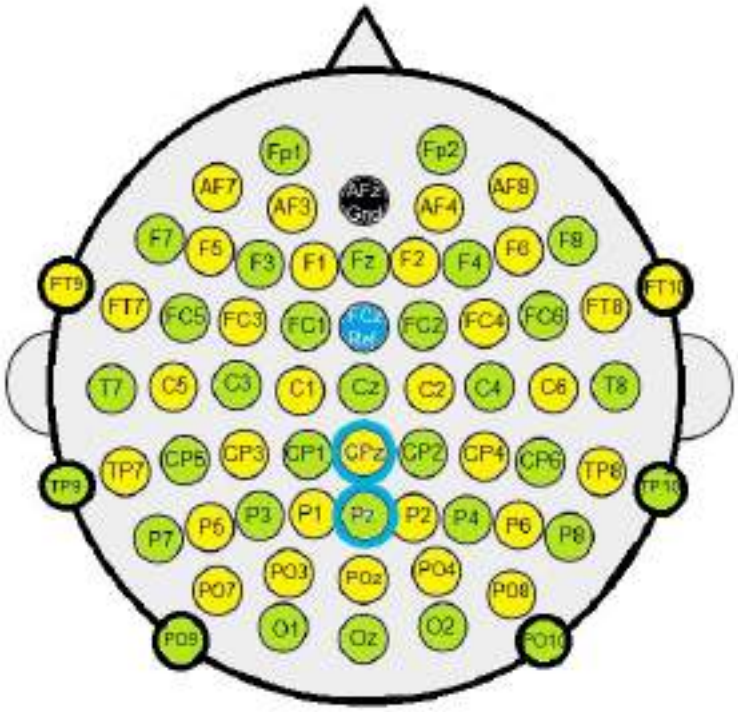

Figure 7.5: Electrodes in the 64 channel EEG head cap. Here green color denotes electrode positions for channels 1 to 32, yellow color for channels 33 to 64, blue electrode depicts the reference channel and black depicts the ground electrode. Also, frontal channels F3 and F4 are representative of the frontal alpha asymmetry index (FAA).

used to detect electro-oculographic (EOG) artifacts. The EEG raw data was synchronized with the experiments. The signals were then epoched into 1 second segments and filtered via a least squares filter. The segments corresponding to bad EEG data were rejected by setting a threshold for the peak-to-peak amplitude. To further reject artifacts corresponding to eye blinks or muscles, independent component analysis (ICA) was carried out [242]. The relevant independent

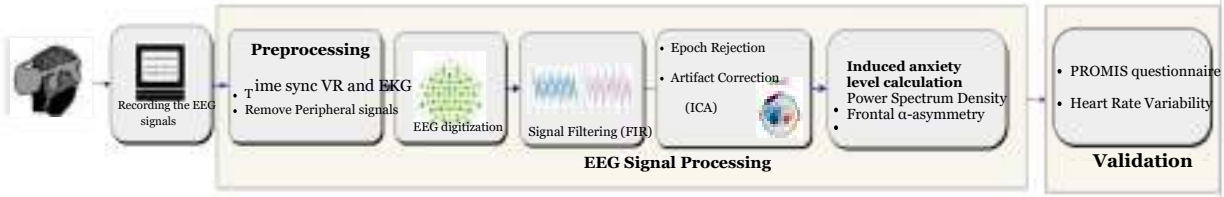

Figure 7.6: EEG analysis pipeline

components capturing anxiety were preserved to compute the power spectrum density (PSD), which was used to extract features (in particular the frontal alpha asymmetry index [238]) to evaluate human anxiety. A detailed work flow pipeline for the EEG data analysis is shown in Figure 7.6. A summary of the major processing steps is highlighted as follows.

• Filtering: The recorded EEG signals from 64 distinct channels of the brain are inherently noisy. Hence, they were processed using a band pass finite impulse response filter (FIR) (eq. 7.1). FIR filters, being stable, are a recommended choice in EEG signal refining [243].

$$f[i] = \sum_{n=0}^{N} n\, z[i-n] \tag{7.1}$$

• Epoch rejection: To reject epochs with bad data, thresholds were defined for peak-to-peak amplitude detection. Variance and voltage changes in signals were also used as features to reject bad sections in the data.

• Artifact rejection: Since the filtered EEG data is a linear combination of cortical oscillations from 64 independent channels, ICA is a rational approach to perform artifact correction [244]. Defining S to be a matrix of the recorded EEG data, ICA estimates a weights matrix W and independent components (ICs) C ($S = W^{-1}C$) to maximize statistical independence among components. The extracted independent components can be categorized into seven distinct classes [245], [246], namely, i) brain components referring to stable EEG signals from the cerebral cortex, ii) eye components generated from the retina, responsible for producing vertical and horizontal eye movements, iii) heart components originating from the cardiac muscles, iv) muscle components from the muscle activities like the neck movements, v) line noise from the external electronic devices, vi) channel noise describing distortion from a single EEG channel and vii) other artifacts that do not qualify as relevant brain components for further analysis. The brain components were the only acceptable independent components for PSD feature extraction. Hence, to identify and eliminate the eye, heart, muscle or externally generated artifacts in the