ML-ASSISTED THERAPEUTICS FOR NEURODEGENERATIVE DISORDERS

BY

ANANT DADU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Roy H. Campbell, Chair
Professor Jimeng Sun
Professor Minh N. Do
Dr. Mike Nalls, National Institutes of Health
Dr. Faraz Faghri, National Institutes of Health

ABSTRACT

Neurodegenerative disorders (NDDs) are a significant public health issue, affecting 50 million people worldwide every year. The complexity of NDDs hinders progress in the development of prevention and disease-modifying therapies. Despite numerous clinical trials, the success rate for treating the condition remains less than 1%, with many trials failing at the late stage leading to significant financial burden and negative outcomes. Challenges presented by NDDs include disease heterogeneity, overlapping clinical syndromes, a long asymptomatic phase, and incomplete understanding of disease mechanisms. A more systematic and efficient approach to the causes and diagnosis of these diseases is needed to accelerate the growth of effective treatments and ultimately improve health outcomes.

In the current research landscape, there has been a remarkable upsurge in real-world datasets dedicated to NDDs, characterized by a significant expansion in both sample size and the inclusion of diverse data modalities. Leveraging machine learning techniques to analyze this data presents an exciting opportunity to address challenges presented by NDDs. We have shown that a machine learning algorithm can delineate subgroups within Parkinson's disease by discovering hidden patterns from multi-modal symptomatic data in an unbiased way. Given the longitudinal nature of NDDs, we illustrated the use of longitudinal dimensional reduction approach to identify underlying trajectory patterns within large biomedical datasets. We demonstrated that disease probability scores obtained by exposing brain imaging and genomics data to machine learning tools are useful for risk stratification, prognosis prediction, and monitoring disease progression. Our multi-modal approach on large aggregates of real-world data, along with the contribution of our interactive data-driven web applications, leads to a substantial enhancement in transparency, reproducibility, and accessibility.

We anticipate that this dissertation will have a transformative impact on industry and academia by advocating for and enabling data-driven methodologies to enhance medical research. Our comprehensive evaluation and open-source deployment of research results should reduce the friction between basic science research and its practical implementation in clinical settings or drug development processes. As research evolves and produces more complex datasets, we believe that the use of computational tools will become more prevalent in the field. Research outputs of this work can serve as a reference for future research in this area, as it showcases the potential of machine learning to assist medical research for neurodegenerative disorders.

To my family, for their love and support.

ACKNOWLEDGMENTS

This dissertation is not possible without the support and guidance of many people.

First and foremost, I am deeply grateful to my advisor, Professor Roy H. Campbell, whose mentorship has profoundly shaped my academic journey. He has not only taught me to think with a long-term vision but also guided me in asking the right questions and effectively presenting my ideas with passion. His unwavering support throughout my PhD has been invaluable. Roy provided me with the freedom to explore my interests while offering guidance on aligning ideas and gaining perspective. Moreover, he consistently presented numerous opportunities for each of my projects during my doctoral studies.

Dr. Faraz Faghri has been an invaluable mentor and guide throughout my journey, serving as an advisor alongside Roy. His enthusiasm for new ideas and technological innovations has been truly inspiring. Under his guidance, I have learned to think critically and ask the right questions, always understanding the purpose behind every endeavor. Faraz's motivation for user adoption has driven me to see our projects through to their ultimate goals, ensuring people use our work. He has significantly enhanced my presentation skills, enabling me to effectively communicate complex concepts in a clear and concise manner. Moreover, his mentorship has contributed to my overall personal growth as both a researcher and an individual.

Dr. Mike Nalls has played an instrumental role making this work possible. His passion for advancing treatments for neurodegeneration has been an immense source of motivation for me. Mike's dedication to open science in biomedical research has instilled in me a profound understanding of its importance. From him, I have learned the significance of approaching complex tasks by focusing on simple solutions. Finally, this research would not have been possible without the support and continuous guidance from Dr. Andrew Singleton.

Thanks to my amazing committee members, Professor Jimeng Sun and Professor Minh N. Do. I learned a lot as a teaching assistant at Jimeng's deep learning for healthcare course. Both Minh and Jimeng provided me with critical feedback on my work and encouraged me to evaluate the practical implications of my research.

I feel incredibly fortunate to have outstanding friends and mentors at the National Institute of Health, including Sara Bandres Ciga, Mary Makarious, Jeff Kim, Marianita Santiana, Mathew Koretsky, Andy Qi, Michael Ward, Mark Cookson, and Lana Sargent. Additionally, at Data Tecnica International, I am grateful for the guidance and support of Michael Ta,

Kristin Levine, Nicholas Johnson, Hampton Leonard, and Hirotaka Iwaki. I learned amazing things in our weekly data science team meetings.

I would like to express my sincere gratitude to my colleagues and friends at the University of Illinois at Urbana Champaign: Harshit, Amit, Vedaanta, and Priyank. I am also grateful to Rachneet and Vipul for their contributions and collaboration. Furthermore, I extend my thanks to my collaborators at Invicro, including Brian Avants, Jacob Hesterman, and Roger Gunn, for their valuable insights and cooperation.

Many thanks to the staff of the Computer Science department, including but not limited to: Viveka Perera Kudaligama, Jennifer Comstock and Kara MacGregor for their help at every point during my PhD. At NIH, thanks to Jeysson for addressing all my computer related concerns and Mackenzie, Charlotte who went across all the complicated paperwork required.

I am grateful to receive financial support from the National Institute of Health, Center for Alzheimer's and Related Dementias, and Data Tecnica International. They also provided access to real world healthcare data sets and cloud computing services. I would like to extend my heartfelt thanks to the participants, patients, and their families who contributed to this research, as they played a key role in the completion of this dissertation.

Finally, this dissertation is dedicated to my parents, Alok and Sapna Dadu, my brother, Shrey Dadu, and my sister, Vidushi Dadu who supported me throughout my PhD journey. My parents provided emotional support by keeping belief, and pride in me throughout my journey. Shrey has always been my go-to person. Vidushi, with her forceful advice and willingness to push me out of my comfort zone, has helped me overcome numerous obstacles. They all have provided a listening ear, offering reassurance and comfort during the most challenging times. Their presence in my life has been truly priceless.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 Background 1
1.2 Thesis Statement 3
1.3 Dissertation Foundations 3
1.4 Related Works 6
1.5 Dissertation Contributions 8
1.6 Dissertation Outline 10

CHAPTER 2 PARKINSON'S DISEASE PROGRESSION BASED SUBTYPES . . 12
2.1 Introduction 12
2.2 Methods 14
2.3 Results 22
2.4 Discussion 34

CHAPTER 3 APPLICATION OF ALIGNED-UMAP TO LONGITUDINAL BIOMEDICAL STUDIES 46
3.1 Introduction 46
3.2 Methods 49
3.3 Results 51
3.4 Discussion 57
3.5 Future Work 59

CHAPTER 4 PREDICTION, PROGNOSIS AND MONITORING OF NEURODEGENERATION AT BIOBANK-SCALE 62
4.1 Introduction 63
4.2 Research in context 65
4.3 Methods 66
4.4 Results 69
4.5 Discussion 75

CHAPTER 5 TRANSPARENT, REPRODUCIBLE AND ACCESSIBLE SCIENCE 78
5.1 Introduction 78
5.2 Methods 78
5.3 Results 80
5.4 Discussion 85

CHAPTER 6 CONCLUSIONS 89
6.1 Introduction 89
6.2 Contributions 90
6.3 Future Work 92
6.4 Broader Impacts 98

REFERENCES 99

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

A disease is a deviation from an organism's normal structural and functional behavior. These behaviors are called signs and symptoms of a disease. Diseases may often be classified by cause, pathogenesis, or symptoms. Today, it is preferred to classify disease based on cause if it is known. The disease classification scheme gets updated iteratively as more detailed biological data is collected. Forming the diagnostic criteria for a disease is the first step in the process of drug development.

Medicine is the science of restoring or preserving physical condition by managing, preventing, and treating the disease using artifacts such as drugs or surgical appliances of operations. Diseases are associated with fundamental processes that cause particular observations or phenotypes. These processes, more often, are defects in molecular and cellular processes that constitute the triggers of specific pathology and are referred to as mechanisms of disease. Understanding disease mechanisms is crucial to discovering and providing the proper treatment to improve the healthcare of suffering patients.

Following disease classification, medical research is broadly classified into basic science and translational research. Basic science focuses on the discovery of clinical therapeutic candidates. Translational research is taking discovery from the laboratory into the clinic, where it can ultimately help people. The procedures used in basic science research vary based on the disease complexity or nature of the disease. Research for illnesses that are caused primarily due to environmental effects is treated differently compared to genetic diseases such as cancer or neurodegenerative disorders. Drugs addressing targets supported by human genetic evidence are more likely to progress through clinical trials [1]. Two-thirds of FDA-approved drugs in 2021 have shown genetic evidence [2]. Here we focus on the pipeline for genetics-based drug discovery, focusing on the steps involving biological data.

- **Basic science research** includes:
    - Genome-wide association study (GWAS) is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait.
    - Functional Genomics studies understanding gene functions and their interaction with other genes or proteins that result in a particular trait. GWAS studies cannot distinguish between causal and non-causal genes, but Functional genomics usually

  tests the “causal role” of the function. It also includes understanding genes’ interaction with environmental factors and cell biology.
  - Once the potential therapeutic target is identified, drugs are being developed that can interfere with the therapeutic target to correct defects in the pathology. The drugs are validated in screenable assays and animal models before clinical trials.

- **Translational research**: Once a drug is developed and has shown promise in preclinical research; it goes through clinical trials. First, a drug is tested on healthy people for safety and toxicity (Phase 1 trial). Then, drugs are given to patients in a controlled environment to test for efficacy and side effects (Phase 2 and Phase 3 trials). If approved, the drug is translated into practice and then applied to the community for the actual benefit of society.

Machine learning is a branch of Artificial Intelligence (AI) that leverages vast amounts of data to learn hidden patterns to improve performance on a given set of tasks. The growth of high-quality large datasets and the emergence of neural network-based specialized algorithms led to the prevalence of AI in society [3, 4, 5, 6]. In recent years, AI has achieved tremendous success in medicinal applications outperforming expert humans on various tasks. For instance, pneumonia diagnosis using chest X-rays [7, 8] or skin cancer detection using skin images [9]. The success of machine learning algorithms is attributed to their ability to deal with tasks whose complexity challenges the mental capacity of a human brain to solve. The human body is a very complex system; biologists or clinicians can only analyze a handful of observations. Another exciting area of research, multi-modal machine learning, aims to build models that can process and relate information from multiple modalities [10]. Therefore, for a disease, we believe that machine learning can be a handy tool in medicine and can improve medicinal science significantly.

Neurodegenerative disease is an overarching term for various progressive conditions that affect the nervous system’s neurons, resulting in the damage and death of brain cells. This abnormality causes deficiency related to cognitive and movement skills in the affected people. We identified four critical characteristics of neurodegenerative disorders that contribute to their complexity and can be addressed using data-driven studies: (1) **disease heterogeneity**, (2) **overlapping clinical syndromes**, (3) **long asymptomatic phase**, and (4) **understanding disease mechanisms**. Disease heterogeneity refers to a medical condition that shows substantial phenotypic variability or variability in the root cause of the disease patients. Data-driven studies have identified significant variation in clinical manifestations by age at onset, rate of progression, associated treatment complications, and the occurrence and constellation of the motor/non-motor features for Alzheimer’s and Parkinson’s disease.

The phenotypic heterogeneity within the diseased population poses a significant challenge for clinical care and clinical trial design. The other important aspect of neurodegenerative disorders is the overlap in clinical symptoms and risk factors [11]. For example, cognitive deficits occur not only in Alzheimer's disease (AD) but also in vascular dementia (VD), frontotemporal dementia (FTD), mixed dementia, and dementia with Lewy bodies (LBD). Similarly, the motor system is affected by amyotrophic lateral sclerosis (ALS), Huntington's disease (HD), Parkinson's disease (PD), and spinocerebellar ataxias (SCAs). In addition, aging is a significant factor common across NDDs. Recognizing shared pathways across neurodegenerative disorders can lead us to potential therapeutic targets while reducing the chances of related disorders at older ages [12]. Research shows that NDDs generally start decades before symptom onset [13, 14]. Disease-modifying therapies are most effective during this early stage of the disease, making early intervention a crucial factor [15]. Finally, even though it is known that NDDs pathogenesis involves a complex concert of events and molecular players, current strategies for therapeutic targets primarily focus on hypothesis-driven approaches where single, predefined cellular readouts or genetic variants are targeted.

1.2 THESIS STATEMENT

Machine learning enhances therapeutic discovery of neurodegenerative disorders. Multimodal analysis of real-world biological datasets, including, multi-omics, clinical symptoms, and cellular and brain imaging were used to demonstrate utility of machine learning.

1.3 DISSERTATION FOUNDATIONS

The objective of our thesis, entitled "ML-Assisted therapeutics for Neurodegenerative Disorders" is to investigate the potential of Machine Learning to overcome the obstacles that lead to the failure of clinical trials in NDDs. This thesis rests on two fundamental pillars, which are: (1) the current body of research and observations in the field of NDDs, (2) availability of datasets and machine learning algorithms. In the following sections, we provide a more detailed explanation of each of these components.

1.3.1 Current body of research and observations in the field of NDDs

NDDs poses three significant and enduring challenges that hinder the development of effective drugs. Here, we discuss these challenges and how addressing them can lead to better clinical outcomes.

- **Within disease subgroups**: The phenotypic heterogeneity that exists within the neurodegenerative diseased population poses a major challenge for clinical care and clinical trial design. A clinical trial has to be suitably tailored to account for inter-individual variability, and as a consequence, trials are either large, long, expensive, and/or only empowered to see large effects. This problem becomes particularly burdensome as we move increasingly towards early-stage trials when therapeutic interventions are likely to be most effective. To that effect, defining subcategories of NDDs and the ability to predict even a proportion of the disease course has the potential to significantly improve cohort selection, inform clinical trial design, reduce the cost of clinical trials, and increase the ability of such trials to detect treatment effects.

- **Overlapping signs and symptoms**: All NDDs have different clinical entities concerning affected brain regions, progression, onset age, and behavior traits. However, they share lot of similarities in clinical signs and symptoms. The distinction between motor and cognitive skill deficiencies across NDDs is unclear. This leads to a high rate of misdiagnosis and impacting clinical outcomes. To improve diagnostic accuracy, it is crucial to move beyond signs and symptoms and incorporate more sophisticated data modalities, such as brain imaging and genome sequencing. Further, these commonalities suggest that there might be an overlap between many of these disorders at the gene level. It can provide insight into better drug targets and lead to fundamental mechanisms behind neurodegenerative diseases. Understanding the differences/similarities between these disorders can provide us with better recruitment criteria for clinical trials.

- **Long asymptomatic phase**: Both Alzheimer's disease (AD) and Parkinson's disease (PD) pathology starts decades before clinical symptoms appear. This long preclinical phase of neurodegenerative disorders (NDDs) offers an opportunity for early intervention with disease-modifying therapies, where chances of successful treatments are high [16]. However, the lack of biomarkers for early diagnosis and disease progression monitoring remains a significant obstacle to achieving this goal. Genomics and imaging based predictive models can give early disease detection and precision medicine tools [17].

- **Insufficient disease understanding**: As the need for treatment of complex disorders such as NDDs grows, because of ageing population, we will need ways to quickly model them using accurate phenotypes, validate them, and improve our understanding of the disease's genetic architecture. Data-driven methodologies can help in providing a better understanding of disease pathophysiology. Drug targets that show evidence of potential treatment have a higher chance of moving to clinical trials and potentially disease treatment [1, 2]. The availability of large-scale biomedical databases and research resources, such as the UK biobank, containing in-depth genetic and health information provides ample opportunities for data-driven analysis. In addition, innovations in biological tools such as CRISPR and iPSC allow us to collect datasets in vitro to understand complex diseases better. However, the optimal integration of modern machine learning methods and biological data focusing on NDDs is still lacking. Integrating biological understanding of disease into machine learning models would be better. Moreover, we need systematic and reproducible tools to take the discoveries to translational science. We discuss our ideas to enhance the understanding of disease mechanisms by discovering novel disease risk genes and monitoring changes in cell morphologies.

1.3.2 Availability of large scale datasets and machine learning algorithms

The second pillar of this work is based on the availability of comprehensive and carefully maintained large-scale databases. Neurodegenerative diseases are complex and often affect multiple dimensions of health, as seen in Parkinson's disease, where both cognitive and motor skills are impacted. Additionally, these disorders can begin developing decades before any symptoms arise, making them especially challenging to treat. To address this, multiple study cohorts have been established to gather as much data as possible in order to gain a deeper understanding of these diseases. In addition to large number of sample, we also require large number of features to address multidimensional nature of these disorders. This dissertation builds upon these well-curated cohorts, including Alzheimer's Disease Neuroimaging Initiative, Parkinson's Progression Markers Initiative, Parkinson's Disease Biomarkers Program, UK Biobank, and several multi-cohort data. We analyzed the datasets using different modalities such as clinical signs, Magnetic Resonance Imaging, whole genome sequence, and blood biomarkers. Table 1.1 provides a quick overview of the datasets we used in this work. It is impossible to analyze or interpret this high-dimensional data without the use of machine learning. We utilize machine learning techniques, that includes supervised, unsupervised, and semi-supervised learning, to analyze the data and develop models. Unsupervised models are used for detecting subgroups and supervised models were used for risk, prognosis prediction in NDDs.

Table 1.1: Shows a list of all the datasets we used in our analysis from different biological modalities. (NDD: Neurodegenerative disorders, PPMI: Parkinson's Progression Markers Initiative, PDBP: Parkinson's Disease biomarkers discovery, ADNI: Alzheimer's Disease Neuroimaging Initiative, UKBB: UK Biobank, iNDI: iPSC Neurodegenerative Disease Initiative, NIH: National Institutes of Health)

Aspect	**NDD**	**Modalities Involved**	**Participants (count)**	**Study**
Disease Heterogeneity	HC	Signs/Symptoms	154	PPMI
	HC	Signs/Symptoms	115	PDBP
	PD	Signs/Symptoms	294	PPMI
	PD	Signs/Symptoms	263	PDBP
	HC	Signs/Symptoms	180	ADNI
	AD	Signs/Symptoms	254	ADNI
Shared mechanisms	HC	Genomics (SNP variants)	4,027	Multi Cohort
	PD	Genomics (SNP variants)	-	Multi Cohort
	AD	Genomics (SNP variants)	-	Multi Cohort
	ALS	Genomics (SNP variants)	1,052	Multi Cohort
	DLB	Genomics (SNP variants)	2,590	Multi Cohort
	FTD	Genomics (SNP variants)	1,386	Multi Cohort
Disease mechanisms (GWAS)	HC	Genomics (SNP variants) Brain Imaging (MRI)	∼45000	PPMI, PDBP, ADNI, UKBB
	AD	Genomics (SNP variants) Brain Imaging (MRI)	∼3000	PPMI, PDBP, ADNI, UKBB
	PD	Genomics (SNP variants) Brain Imaging (MRI)	∼2000	PPMI, PDBP, ADNI, UKBB
Disease mechanisms (Cell Morphology)	DLB	Cell Microscopy Imaging	15 genes	iNDI NIH
	FTD	Cell Microscopy Imaging	28 genes	iNDI NIH
	AD	Cell Microscopy Imaging	8 genes	iNDI NIH

1.4 RELATED WORKS

Here we discuss related works with respect to each of the challenges presented by NDDs and how previous research has dealt with it.

1.4.1 Disease heterogeneity

- **PD subtypes**: Recently, subtyping studies using a variety of phenotypes across neurodegenerative disorders have gained significant attention. The earlier works have followed a path of clinical observation based on age at onset or categorization based on the most observable features for PD subtypes [18, 19, 20]. This dichotomous separation, while intuitive, does not faithfully represent the clinical features of the disease, which are quantitative, complex, and interrelated. A more realistic representation of the disease and disease course requires a transition to a data-driven, multi-dimensional schema that encapsulates the constellation of interrelated features and allows tracking (and ultimately predicting) change. Previous studies used cluster analysis, a data-driven approach, to define two to three clinical PD subtypes [21, 22, 23, 24]. Depth of phenotypic information and longitudinal assessments in these studies were variable and often limited to certain clinical features and short-term follow-ups.

- **AD subtypes**: Machine learning has successfully defined subtypes showing heterogeneity in affected brain regions and the progression [25, 26]. Progression-based subtypes are also identified by exposing unsupervised machine learning algorithms to cognitive symptoms and signs data [27].
- **ALS subtypes**: Given the importance of clinical heterogeneity within ALS, there has been a considerable effort to develop classification systems for patients over time. Examples include groupings based on family status [28], clinical milestones [29]), neurophysiological measures [30], and diagnostic certainty [31]. Machine learning algorithms have shown efficacy in delineating ALS subtypes [32].

1.4.2 Overlapping clinical syndromes

Although neurodegenerative disorders differ from one another with regard to core clinical pathologies, they have fundamental commonalities at the genomics level as well as symptomatic level [11]. For example, hexanucleotide repeat expansion mutations in intron 1 of C9orf72 causes ALS in some family members and FTD in others [33, 34]. Attempts thus far have focused on analyzing post hoc GWAS summary statistics, understanding shared risk variants and comparing polygenic scores across distinct diseased subjects [35]. There are studies doing pleiotropic analysis of linking easy to obtain phenotypes such as aging, cognitive and shows associations with neurodegenerative disorders [12]. None of the previous studies have focused on utilizing ML based approaches to understand the brain regional similarities between all these neurodegenerative disorders. The accurate diagnosis of NDD types is challenging due to the presence of syndromes with similar phenotypes. For instance, the most accurate method of determining dementia type is through brain biopsy for research purposes.

1.4.3 Disease mechanisms

Gene defects is the fundamental step towards understanding the cause of a disease at the most basic level – the cell and the nucleotide [36]. Genetics has been identified to be a significant contributor for neurodegenerative disorders [37]. The growth of genomic wide sequencing technologies led GWAS studies for the identification of disease etiology for AD [38], PD [39], ALS [40], FTD [41] and LBD [35]. Though these studies have utilized genomic data from thousands of individuals, they are dependent on proxy cases and cannot address heterogeneity within diseased individuals. Recently, GWAS have been applied on more accurate phenotypes determined using machine learning techniques [42, 43]. Unsupervised algorithms (PCA) have been applied to obtain COVID-19 severity measures [42]. ML based phenotypes obtained from color fundus photographs have identified 93 novel loci for glaucoma [43]. Despite the identification of any risk genes, the function of many genes and gene regulatory elements remains poorly characterized, which limits our ability to apply these insights towards disease treatment [44]. It becomes particularly challenging for NDDs as we could not monitor the diseased cells due to their death, decades before symptom onset [45]. The advent of CRISPR, and iPSC allows for monitoring neuron changes under wild type and mutated conditions [46]. With respect to phenotypes, high throughput screening using confocal microscopy and cell painting allows us to capture morphological changes in disease neurons [47, 48].

1.5 DISSERTATION CONTRIBUTIONS

1.5.1 Publications

1. Dadu, A., Satone, V., Kaur, R. et al. Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Parkinsons Dis. 8, 172 (2022). https://doi.org/10.1038/s41531-022-00439-z* [49]

2. Dadu, Anant, et al. "Application of Aligned-UMAP to longitudinal biomedical studies." Patterns (2022). [50]

3. Prediction, prognosis and monitoring of neurodegeneration at biobank-scale via machine learning and imaging. (Dadu, Anant, et al. *In preparation* for submission at The Lancet Digital Health, 2023)

4. Faghri, Faraz, et al. Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. The Lancet Digital Health 4.5

(2022): e359-e369. [51]

5. Satone, Vipul K., et al. Predicting Alzheimer's disease progression trajectory and clinical subtypes using machine learning." bioRxiv (2019): 792432.

6. Mathew J Koretsky and others, Genetic risk factor clustering within and across neurodegenerative diseases, Brain, 2023 [52]

7. Makarious, Mary B., et al. Multi-modality machine learning predicting Parkinson's disease. npj Parkinson's Disease 8.1 (2022): 35. [53]

8. Makarious, Mary B., et al. GenoML: automated machine learning for genomics. arXiv preprint arXiv:2103.03221 (2021). [54]

9. Reilly, Luke, et al. A fully automated FAIMS-DIA proteomic pipeline for high-throughput characterization of iPSC-derived neurons. bioRxiv (2021): 2021-11. [55]

1.5.2 Web dashboards

(a) Subtypes for Amyotrophic Lateral Sclerosis (ALS): `http://bitly.ws/u5h7`

(b) Subtypes for Alzheimer's disease (AD): `http://bitly.ws/u5h3`

(c) Subtypes for Parkinson's disease (PD): `https://shorturl.at/fkvw3`

(d) Longitudinal Aligned-UMAP: `https://rb.gy/zf2xu`

Our solutions provide a systematic approach to address the challenges presented by neurodegenerative diseases using data-driven analysis. Our work on PD subtypes (1) involved the development of machine learning models that delineate progression-based subgroups using clinical signs and symptoms. This work has been recognized in several medical news articles[1][2][3][4][5]. We have also applied a similar algorithm to tease apart subtypes of AD (5) and contribute to the discovery of ALS subtypes with varying clinical manifestations (4).

Our use of unsupervised learning algorithms to identify hidden patterns from seven longitudinal biomedical datasets has been illustrated in 2, and can also be used to detect batch

[1] https://www.healio.com/news/neurology/20221216/machine-learning-identifies-three-parkinsons-subtypes-to-better-track-disease-progression

[2] https://newslanes.com/2022/12/16/machine-learning-identifies-three-parkinsons-subtypes-to-better/

[3] https://parkinsonsnewstoday.com/news/algorithm-able-predict-5-year-rate-parkinsons-progression/

[4] https://www.neurologyadvisor.com/topics/movement-disorders/machine-learning-models-identify-subtypes-parkinson-disease/

[5] https://neuroderm.com/living-with-parkinson-s/info-center/machine-learning/

effects in biological experiments (9). We have also made significant contributions towards increasing interpretability of supervised models, providing insights into factors contributing to Parkinson's disease (7, 8). Analyzing whole genome sequencing data across neurodegenerative disorders yielded genomic-based clusters using an unsupervised learning approach (6).

We have also applied supervised learning techniques to brain MRI data, developing early detection tools for dementia and PD (3). When combined with genomic data, it provides insights into the overlapping nature of neurodegenerative disorders (3). Finally, we have put significant effort into increasing transparency, reproducibility, and accessibility of the research output of this thesis (a, b, c, d). Refer to Table 1.2 for the summary of dissertation contributions.

Table 1.2: Shows the summary of a list of challenges and the contributions of the dissertation.

Challenges	**Contributions**			
Within disease subgroups	C1: Identification of progression based PD subgroups C2: Prediction of subgroups within 1-2 years of diagnosis using 20 features C3: Replication in an independent cohort			
Longitudinal pattern discovery	C4: Visualize hidden patterns detecting batch effects, clusters			
Early diagnosis	C5: Early detection for both Parkinson's and Alzheimer using brain imaging C6: Replication early detection in an independent cohort			
Overlapping diseases	C7: Added insights into AD and PD overlaps using imaging C8: Contributed to identification of common genes across NDDs			
Disease understanding	C9: Top gene expression profiles associated to PD C10: Top clinical features associated to PD, AD and ALS subgroups			
Open-source tools	C11: Develop prototypical websites for exploring PD, AD and ALS subgroups C12: Developed website showing patterns in longitudinal data			
Aim	**Title**	**Methods**	**Post prelim**	**Contributions**
Chapter 2: Disease subgroups	Subtype identification for PD	Unsupervised ML, Latent space, Clustering	Published npj Parkinson's	C1,C2,C3,C11
Chapter 3: Disease subgroups	Applications of Aligned-UMAP to biomedical studies	Unsupervised ML, Longitudinal data, Webtool	Published Cell Patterns	C4,C12
Chapter 4: Early diagnosis & overlaps	Prediction and prognosis of neurodegeneration at biobank-scale	Supervised ML, Mixed effects model, Image analysis	New results, Replication In preparation	C5,C7,C8
Chapter 5: Open-source tools	Transparency, reproducibility and accessibility of our research	Feature interpretation, SHAP values, Webtool	MVPs deployed	C11,C12
Disease subgroups	Subtype identification for ALS	Unsupervised ML, Latent space, Clustering	Published Lancet digital	C10,C12
Disease subgroups	Subtype identification for AD	Unsupervised ML, Latent space, Clustering	In submission Alz & Dementia	C10,C12
Overlaps	Genetic clustering in neurodegeneration	Unsupervised ML	Published Brain	C8
Disease mechanisms	Multi-modality machine learning predicting PD	Ensemble ML	Published npj Parkinson's	C9
Disease mechanisms	Automated proteomic pipeline for iPSC- neurons	Unsupervised ML, Longitudinal data	In submission	C4
Methods	GenoML: automated ML for genomics	ML applied to genomics	In submission	C9

1.6 DISSERTATION OUTLINE

This dissertation started with analysis of signs and symptoms for neurodegenerative disorders, followed by solutions to address more complex modalities such as genomics and brain Magnetic Resonance Imaging (MRI) data. It is divided into multiple chapters and a brief description below:

- **Chapter 2** addresses the challenge of *within disease heterogeneity* for Parkinson's disease (PD). We discusses our work on discovery and prediction of progression based subgroups to address Parkinson's disease heterogeneity. The in-depth analysis of these subgroups is conducted using biospecimen biomarkers and genetic data modalities.

- **Chapter 3** presents a deep insight into the application of Aligned-UMAP, a dimensionality reduction tool specifically designed for longitudinal datasets, in the biomedical domain. We applied the algorithm on seven high-dimensional longitudinal biomedical datasets. The datasets comprises of brain imaging, clinical signs and symptoms and multi-omics data modalities. The approach yields interesting patterns that holds the potential of identifying disease clusters, sub-structures, and outliers, detecting batch effects, and quality control measures to perform reliable and accurate downstream analyses.

- **Chapter 4** highlights the potential of quantitative markers generated using machine learning techniques for Alzheimer's disease and related dementias (ADRD) and Parkinson's disease (PD). We show that disease probability scores obtained from brain MRI features are useful for risk stratification, prognosis prediction, and monitoring disease progression. We anticipate that this approach represents a step towards developing machine learning tools for *early intervention* and discover *overlapping characteristics* of NDDs.

- **Chapter 5** discusses about the role of transparent, reproducible and accessible research. We highlights the features of the web dashboards we developed as part of this dissertation. For supervised model, the feature interpretability using SHAP values and model perturbation analysis using what if analysis were incorporated in the cloud deployed web platforms. It includes lower dimensional space exploration dashboards for unsupervised models.

- Finally, in **Chapter 6** we conclude with the summary of contributions and future prospects of this dissertation.

CHAPTER 2: PARKINSON'S DISEASE PROGRESSION BASED SUBTYPES

The clinical manifestations of Parkinson's disease (PD) are characterized by heterogeneity in age at onset, disease duration, rate of progression, and the constellation of motor versus non-motor features. There is an unmet need for the characterization of distinct disease subtypes as well as improved, individualized predictions of the disease course. We used unsupervised and supervised machine learning methods on comprehensive, longitudinal clinical data from the Parkinson's Disease Progression Marker Initiative (n = 294 cases) to identify patient subtypes and to predict disease progression. The resulting models were validated in an independent, clinically well-characterized cohort from the Parkinson's Disease Biomarker Program (n = 263 cases). Our analysis distinguished three distinct disease subtypes with highly predictable progression rates, corresponding to slow, moderate, and fast disease progression. We achieved highly accurate projections of disease progression 5 years after initial diagnosis with an average area under the curve (AUC) of 0.92 (95% CI: 0.95 ± 0.01) for the slower progressing group (PDvec1), 0.87 ± 0.03 for moderate progressors, and 0.95 ± 0.02 for the fast-progressing group (PDvec3). We identified serum neurofilament light as a significant indicator of fast disease progression among other key biomarkers of interest. We replicated these findings in an independent cohort, released the analytical code, and developed models in an open science manner. Our data-driven study provides insights to deconstruct PD heterogeneity. This approach could have immediate implications for clinical trials by improving the detection of significant clinical outcomes. We anticipate that machine learning models will improve patient counseling, clinical trial design, and ultimately individualized patient care.

2.1 INTRODUCTION

Parkinson's disease (PD) is a complex, age-related neurodegenerative disease that is defined by a combination of core diagnostic features, including bradykinesia, rigidity, tremor, and postural instability [56, 57]. Substantial phenotypic heterogeneity is well recognized within the disease, complicating the design and interpretation of clinical trials, and limiting patients' counseling about their prognosis. The clinical manifestations of PD vary by age at onset, rate of progression, associated treatment complications, as well as the occurrence and constellation of motor/nonmotor features.

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

Attempts thus far at the characterization of disease subtypes have followed a path of clinical observation based on age at onset or categorization based on the most observable features [18]. Thus, the disease is often separated into early-onset versus late-onset disease, slowly-progressing "benign" versus fast-progressing "malignant" subtypes, PD with or without dementia, or based on the most prominent clinical signs into a tremor-dominant versus a postural instability with gait disorder subtype [19, 20]. This dichotomous separation, while intuitive, does not faithfully represent the clinical features of the disease, which are quantitative, complex, and interrelated. A more realistic representation of the disease and disease course requires a transition to a data-driven, multi-dimensional schema that encapsulates the constellation of interrelated features and allows tracking (and ultimately predicting) change [32, 58].

Previous studies used cluster analysis, a data-driven approach, to define two to three clinical PD subtypes [21, 22, 23, 24]. Depth of phenotypic information and longitudinal assessments in these studies were variable and often limited to certain clinical features and short-term follow-ups. Moreover, many previous studies were limited by insufficient methods to capture longitudinal changes over multiple assessment visits. To this date, none of the previous approaches to PD clustering were replicated in an independent cohort with transparent code and analysis.

We have previously used multi-modal data to produce a highly accurate disease status classification and to distinguish PD-mimic syndromes from PD [59]. These efforts demonstrated the utility of data-driven approaches in the dissection of complex traits and have also led us to the next logical step in disease prediction: supplementing the prediction of whether a person has or will have PD also to include a prediction of the timing and directionality of the course of their disease.

Here, we describe our work on delineating and predicting the clinical progression of PD and for a workflow of the analysis, please refer to Figure 2.1. The first stage of this effort requires creating a multi–dimensional space that captures the disease's features and the progression rate of these features (i.e., velocity). Rather than creating a space based on a priori concepts of differential symptoms, we used data dimensionality reduction methods on the complex

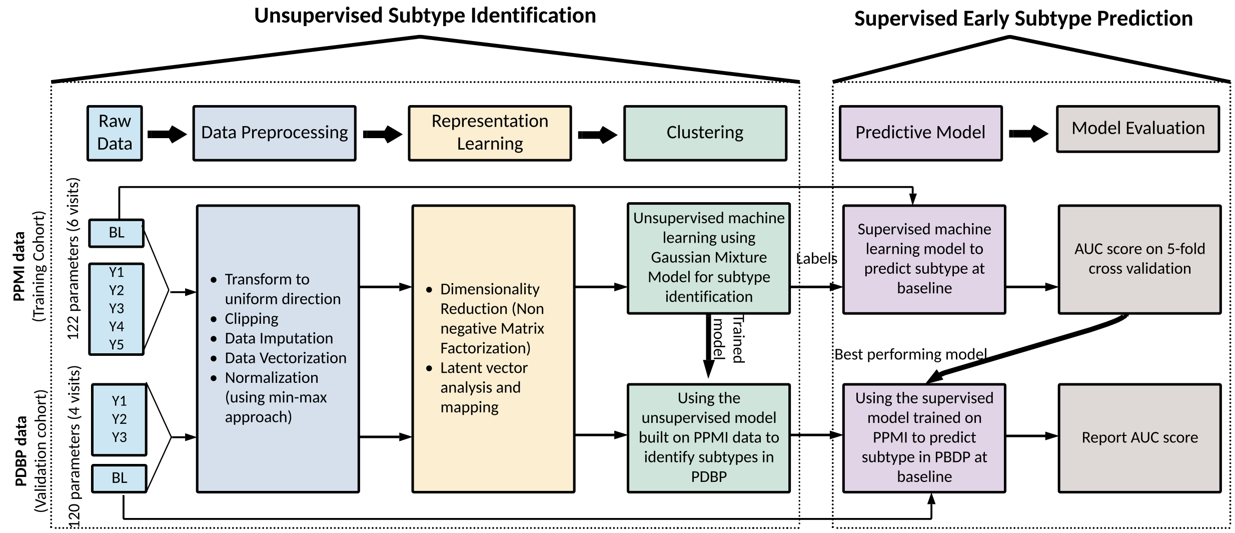

Unsupervised Subtype Identification
Supervised Early Subtype Prediction
Raw Data
Data Preprocessing
Representation Learning
Clustering
Predictive Model
Model Evaluation
PPMI data
(Training Cohort)
122 parameters (6 visits)
BL
Y1
Y2
Y3
Y4
Y5
PDBP data
(Validation cohort)
120 parameters (4 visits)
Y1
Y2
Y3
BL
• Transform to uniform direction
• Clipping
• Data Imputation
• Data Vectorization
• Normalization (using min-max approach)
• Dimensionality Reduction (Non negative Matrix Factorization)
• Latent vector analysis and mapping
Unsupervised machine learning using Gaussian Mixture Model for subtype identification
Trained model
Using the unsupervised model built on PPMI data to identify subtypes in PDBP
Labels
Supervised machine learning model to predict subtype at baseline
AUC score on 5-fold cross validation
Best performing model
Using the supervised model trained on PPMI to predict subtype in PBDP at baseline
Report AUC score

Figure 2.1: Workflow of analysis and model development.

clinical features observed 60 months after initial diagnosis to create a meaningful spatial representation of each patient's status at this time point. After creating this space, we used unsupervised clustering to determine whether there were clear subtypes of disease within this space. This effort identified three distinct clinical subtypes corresponding to three groups of patients progressing at varying velocities (i.e., slow, moderate, and fast progressors). These subtypes were validated and replicated in an independent cohort. Following the successful creation of disease subtypes within a progression space, we created a baseline predictor that accurately predicted an individual patient's clinical group membership five years later. Further, we examined the predictive capability of biospecimen biomarkers at baseline and the genetic information in identifying the subtypes. Our work highlights the utility of machine learning as an ancillary diagnostic tool to identify disease subtypes and project individualized progression rates based on model predictions.

2.2 METHODS

2.2.1 Study design and participants

This study included clinical data from the Parkinson's Progression Marker Initiative (PPMI, http://www.ppmi-info.org/) and the Parkinson's Disease Biomarkers Program (PDBP, https://pdbp.ninds.nih.gov/). Both cohort's data went through triage for missing data, 60-month assessment (36-month in PDBP), and comprehensive phenotype collection. Only data from participants with 60 months of follow-up for PPMI and 36 months for PDBP were included in the study. Overall, in the PPMI (n = 294 PD cases including 99 (34%) female; 154 controls including 58 (38%) female), and in the PDBP (n = 263 PD cases including 112 (43%) female; 115 controls including 64 (56%) female) passed the triage. The PPMI average age at the screening of PD cases was 61 ± 9.7 years and 60.3 ± 11 years for controls. The PDBP average age of PD cases was 64.3 ± 8.6 years and 63.6 ± 9.5 years for controls. The PPMI data also included 28 patients with other enrollments (10 PRODROMA; 8 GENPD; 6 GENUN; 3 SWEDD; 1 REGPD), which were excluded. The PPMI and PDBP cohorts consist of observational data from comprehensively characterized PD patients and matched controls. All PD patients fulfilled the UK Brain Bank Criteria [60]. PD subjects enrolled in PPMI were drug naïve (i.e., not much treated with dopaminergic medications) for at least 2-3 years after enrollment. Being drug naïve is beneficial as we propose to build a disease progression tool for PD subtypes during early stages without complications from pharmacological interventions. Control subjects had no clinical signs suggestive of parkinsonism, no evidence of cognitive impairment, and no first-degree relative diagnosed with PD. Age and MDS-UPDRS Part III (objective motor symptom examination by a trained neurologist) distribution of cohorts at baseline were investigated using Kernel Density Estimation (KDE) to show that these independent cohorts are identically distributed and ensure the integrity of replication and validation (Figure 2.2, Table 2.1). Each contributing study abided by the institutional review boards' ethics guidelines. All participants gave informed consent for inclusion in their initial cohorts and subsequent studies. Figure 2.1 provides an overview of the analyses and study design.

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

PPMI vs PDBP (after 3 years)	**t-value (95% CI)**	**p-value (95% CI)**
Age in 2019	-0.41	0.68
MDS UPDRS PartIII	0.29	0.77

2.2.2 Dataset construction

The discovery and replication cohorts include visit data collected every 12 months starting from baseline to 60 months (36 months for PDBP) follow-up. In PPMI, visits at the 6 and 9-month time points from baseline were excluded in our analysis due to the high data missingness rate (50%).

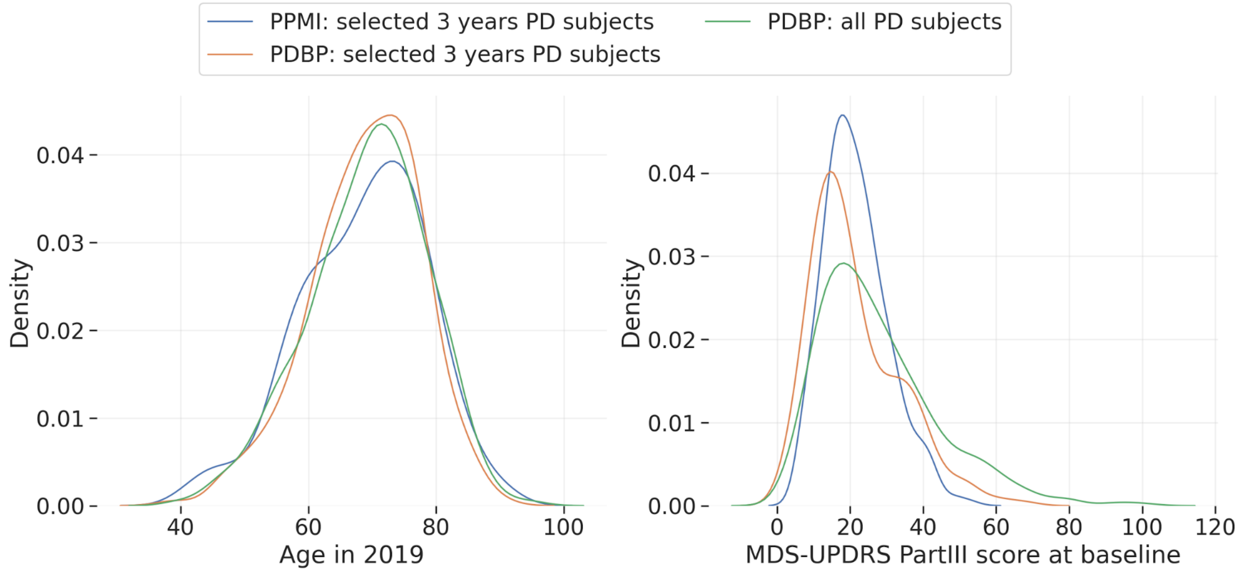


Figure 2.2: Kernel Density Estimation (KDE) analysis of Age and MDS-UPDRS Part III (objective motor symptom examination by a trained neurologist) in PPMI and PDBP cohorts. (a) shows the density of Parkinson's participant's age in the 3-years PPMI, PDBP, and 3-years PDBP datasets, and (b) shows the distribution of Parkinson's participant's MDS-UPDRS Part III at baseline in the 3-years PPMI, PDBP, and 3-years PDBP datasets. The three density functions in both figures are similar showing the validity of statistical replication.

For each cohort, a comprehensive and shared set of longitudinally collected common clinical data elements were selected for analysis. Overall, 122 clinical features were available across six visits for PPMI and 120 features across four visits for PDBP. We used the following features for the subtype identification stage:

- International Parkinson's disease and Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part I, Part II, and Part III [61]
- Cranial Nerve Examination (CN I-XII)
- Montreal Cognitive Assessment [62]
- Hopkins Verbal Learning Test [63]
- Semantic Fluency test [64]
- WAIS-III Letter-Number Sequencing Test [65]
- Judgment of Line Orientation Test [66]

- Symbol Digit Modalities Test
- SCOPA-AUT [67]
- State-Trait Anxiety Inventory for Adults [68]
- Geriatric Depression Scale [69]
- Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease [70]
- REM-Sleep Behavior Disorder Screening Questionnaire [71]
- Epworth Sleepiness Scale [72].

In addition to these clinical measurements, biological and genetic–based features were included in the baseline subtype interpretation and the subtype prediction. These additional features include genotypes using imputed Illumina NeuroX array, vital signs, serum, CSF, and urine measurements. For genotyping data, we used the variants mapped to human genome build 38 (hg38) genotyping from unrelated European ancestry imputed genotype data passing standard QC metrics used to construct the genetic risk score (GRS) [39, 53, 54]. The values indicate the number of copies of the minor allele of each variant for each subject. We used these values as categorical features. Patient characteristics include height, weight, blood pressure, and demographic details. For biological biomarkers, we assessed alpha–synuclein, total tau protein, β–amyloid 1–42 (Aβ42), phospho–tau181 (p–Tau181) in CSF, serum neurofilament light (NfL), and urine levels of di–22 6–bis (monoacylglycerol) phosphate total in the urine. We studied the longitudinal variation of biomarkers and patients' characteristics measurements across the identified PD subtypes. Furthermore, we investigated the biological measurements' role in discriminating PD subtypes.

2.2.3 Procedures and statistical analysis

The data analysis pipeline for this work was performed in Python (version 3.8) with the support of several open-source libraries (NumPy, pandas, matplotlib, seaborn, plotly, scikit-learn, UMAP, XGBoost, LightGBM, H2O, streamlit). To facilitate replication and expansion of our work, we have made the notebook publicly available on GitHub at `https://github.com/anant-dadu/PDProgressionSubtypes`. The code is part of the supplemental information; it includes the rendered Jupyter notebook with full step-by-step data preprocessing, statistical, and machine learning analysis. For readability, machine learning parameters have been described in the Python Jupyter notebook and not in the text of the paper. Our results are available on an interactive web browser (`https://anant-dadu-pdprogressionsubtypes-streamlit-app-aaah95.streamlitapp.com/`), which allows users to browse the PD progression space. In addition, the browser also includes predictive model interpretations allowing readers to explore feature contributions to model performance. For the streamlit website, we designed a surrogate XGBoost classification model for subtype prediction, which uses a single split with 70% training and 30% test data. The reported subtype classification performance in the manuscript is based on a more stringent nested loop procedure.

2.2.4 Data preprocessing

As the clinical features have varying directionalities, features were transformed to ensure the highest values uniformly represent the worst outcome while the lower value corresponds to greater health. To identify the features not following this pattern, we conducted a two-sample one-tailed t-test (null hypothesis: $\mu_{HC} >= \mu_{PD}$) with cases versus controls as two samples. We transformed those features that showed a p-value of less than 0.05 (5% significance test). We further verified the transformations by manually reviewing the distribution of each feature. We performed feature clipping to minimize the influence of extreme outliers in the data. We limited all the features values between the range given by the second and 98^{th} percentiles. Only a few features had residual missingness that was distributed randomly across the patients at a rate of $< 5\%$. For these features, we performed data imputation using linear interpolation longitudinally (i.e., across visits) for each feature. Then, we transformed the dataset into a mathematically meaningful and naturally interpretable format. To achieve this objective, we a) vectorized and b) normalized all longitudinal data. Specifically, we first vectorized by transforming all observations of a particular parameter in a column vector, then appended all parameters together. We then used the min-max method to normalize the data.

2.2.5 Non-negative matrix factorization

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

2.2.6 Latent space adjustment

The progression of space latent vectors (matrix U) learned by NMF shows some weight sharing among the projected dimensions. This weight sharing can be attributed to the presence of correlation in the data between symptomatologies. To represent the progression space so that each progression indicator shows progression velocity for each symptom, we need to adjust the progression space. We performed transformation on NMF learned progression space by taking the weighted sum of the progression indicators. These weights represent the contribution of each dimension for distinct symptomatology.

Through our use of NMF, we identified progressive features based on motor, cognitive, and sleep-based disturbances. Following this, unsupervised learning via Gaussian Mixture Models (GMM) [73] allowed the data to naturally self-organize into different groups relating to velocity of decline across these three categories, from non-PD controls representing normal aging to PD subtypes. GMM is a variant of mixture models, compared to other methods, the parametrization of a GMM allows it to efficiently capture products of variations in natural phenomena where the data is assumed generated from an independent and identically distributed (i.i.d.) mixture of Gaussian (normal) distributions. The assumption of normal distribution (and therefore, the use of GMM) is often used for population-based cohort phenomenon. We use the Bayesian Information Criterion (BIC) to select the number of PD clusters (subtypes) [74]. The BIC method recovers the true number of components in the asymptotic regime (i.e., much data is available, and we assume that the data was generated i.i.d. from a mixture of Gaussian distributions). To replicate the subtype identification, we applied the GMM model developed in the PPMI data to an independent cohort with varying recruitment strategy and design: the PDBP cohort.

2.2.7 Unsupervised subtype identification

We used dimensionality reduction techniques to develop an interpretable representation of high modality longitudinal data. Dimensionality reduction techniques helped us to build the "progression space" where we can approximate each patient's position relative to both controls and other cases after the 60-month period in one-year intervals. We used the Non-negative Matrix Factorization (NMF) technique to achieve this aim [75, 76]. Alternative methods, such as principal component analysis and independent component analysis, did not perform as well as NMF on longitudinal clinical data due to the non-negative nature of our clinical tests. This process essentially collapses mathematically related parameters into the same multi-dimensional space, mapping similar data points close together.

2.2.8 Supervised early subtype prediction

After identifying progression classes using unsupervised learning, we built predictive models utilizing multiple supervised machine learning methods, including the ensemble learning approach. This method combines multiple learning algorithms to generate a better predictive model than could be obtained using a single learning algorithm [77]. To do this, we used stacking ensembles of three supervised machine learning algorithms (Random forest [78], LightGBM [79], and XGBoost [80]) to predict PD clinical subtypes using the data obtained at the time a neurologist first reviewed the patient as the input (combining baseline and varied time points). This approach outperformed other methods in preliminary testing, such as support vector machines (SVM) and simple lasso-regression models. Besides the predictive performance, we chose an ensemble approach due to the nature of our data and problem: (i) decision trees are intrinsically suited for multiclass problems, while SVM is intrinsically two-class, (ii) they work well with a mixture of numerical, categorical, and various scale features, (iii) they can be used to rank the importance of variables in a classification problem and in a natural way which helps the interpretation of clinical results, and (iv) it also gives us the probability of belonging to a class, which is very helpful when dealing with individual subject progression prediction. We developed three predictive models to predict the patient's progression class after 60 months based on varying input factors: (a) from baseline clinical factors, (b) from baseline and first-year clinical factors, and (c) from biomarkers and genetic measurements.

To validate the effectiveness of our predictive models, we used a nested cross–validation (CV) approach with 5 folds in both inner and outer loops. Specifically, we randomly divided the dataset into five subsamples (outer folds). Each of the subsamples was used as the testing data exactly once, while the remaining (training) data was used for hyperparameter tuning and model training. The hyperparameters were chosen based on their average performance on training data during the inner cross–validation loop. The workflow of the approach is depicted in Algorithm 2.1. The performance of the algorithm was measured by the area under the receiver operating curve (AUC) generated by plotting sensitivity vs. (1 – specificity). We used a macro–average AUC score computed by averaging the metric independently for each class (hence treating all classes equally for predicting fast, moderate, and slow progressing cases). The five results from the multiple iterations were averaged to produce a single estimation of performance across these three classes.

Algorithm 2.1: Model Evaluation Procedure

```
for each iteration i=1,2,..I do
    Divide the dataset into train (80%) and test data (20%) at random;
    Divide the train into $K$ cross-validation subfolds at random;
    for each fold k=1,2,..K // CV loop do
        validation = fold $k$;
        subtrain = all folds other than $k$;
        Train model with every hyperparameter on subtrain;
        Evaluate it on validation;
    end
    Calculate the average metrics score on validation over the $K$ folds for every
      hyperparameter;
    Choose the best hyperparameter setting;
    Train a model with the best hyperparameter on train;
    Evaluate its performance on test;
end
Calculate the mean accuracy over all $I$ iterations on test data;
```

To conclusively validate the algorithm, we also evaluated the performance of the predictive models (trained on the PPMI measurements) on the independent PDBP cohort. To replicate, we trained the supervised model on PPMI latent weights (at baseline) and then used the same model on the PDBP latent weights (at baseline). We show that the predictive models preserve their high accuracy applied to another dataset.

2.2.9 Biomarker based prediction

There were 448 observations in total. We did not include plasma, CSF hemoglobin, and CSF glucosylceramide features because of their high missing data ($> 35\%$). Baseline CSF data were missing for p-tau in 51 patients, for total-tau in 26 patients, for abeta-42 in 20 patients, and alpha syn in 15 patients. Serum Nfl is missing in 22 participants, 17 participants did not have data for DNA GRS scores, and 36 participants had missing APOE status. The overall missing percentage for the above measurements was approximately 5%. We imputed the missing predictor variable data with means for numerical features and used most frequent category for categorical features. The remaining features include demographic information (education year, biological sex, birthdate, race), vital signs (weight, height, blood pressure), and family history (parents' PD status) with no patients having missing data for them. For hg-38 genetic measurements, data was missing for 8 participants. We removed these patients from our study. Finally, we had 440 participants and 39 and 64 features for biomarkers and genetics measurements, respectively. All the genetic features are considered as categorical except DNA GRS. For the combined model, we concatenated both the biospecimen and genetic features. This concatenated vector was used as input for the classification of PD subtypes.

2.3 RESULTS

2.3.1 Clustering vectors of progression

Figure 2.3 shows the result of the mathematical projection of PD progression, called Parkinson's disease progression space detailing normalized progression trajectories of each sample relative to others based on this unsupervised classification system. This space shows the relative progression velocity of each patient in 60 months (i.e., speed and direction). The progression velocity level is divided into three main dimensions: motor, cognitive, and sleep-related disturbances. Movement disorders specialists audited component features to categorize these clinical measures into domains of sleep, motor, and cognition disturbance after identification by the algorithm. Based on latent variables clustered within the Parkinson's progression space, the projected motor dimension was responsible for 63.58% of the explained variance, followed by the sleep dimension (21.81%), and cognitive dimension (14.61%). Motor symptoms are the hallmark of PD progression, with sleep and cognitive decline being, in some cases, elevated past that decline seen in controlled aging. The projected motor dimension significantly contributes towards PD progression; however, sleep and cognition are essential, accounting for 37% variation. Across these trajectories, the unsupervised learning analysis reveals and classifies patients into three main subtypes of PD, relating to rates of disease progression: slow progressors (PDvec1), moderate progressors (PDvec2), and fast progressors (PDvec3). This shows how we can map the clinical features and progression velocity from the point of diagnosis. The components of the motor, cognitive, and sleep dimensions with a description of the latent space used to define the progression space that may aid in interpretability are described in next section.

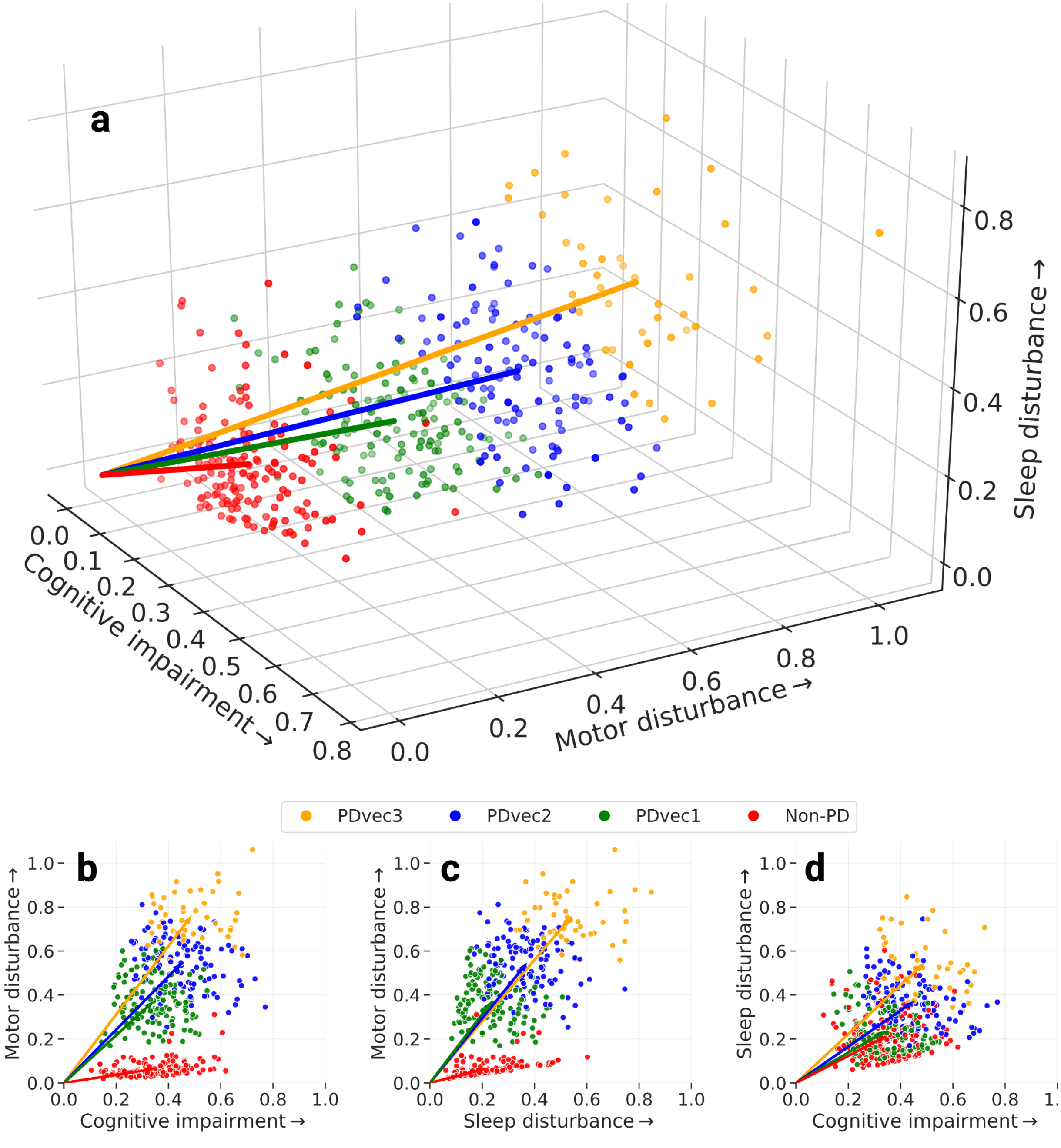


Figure 2.3: Different views of the Parkinson's disease progression space in 5 years with three corresponding projected dimensions cognitive motor and sleep dimensions on a normalized scale.

2.3.2 Interpretation of the Latent Space

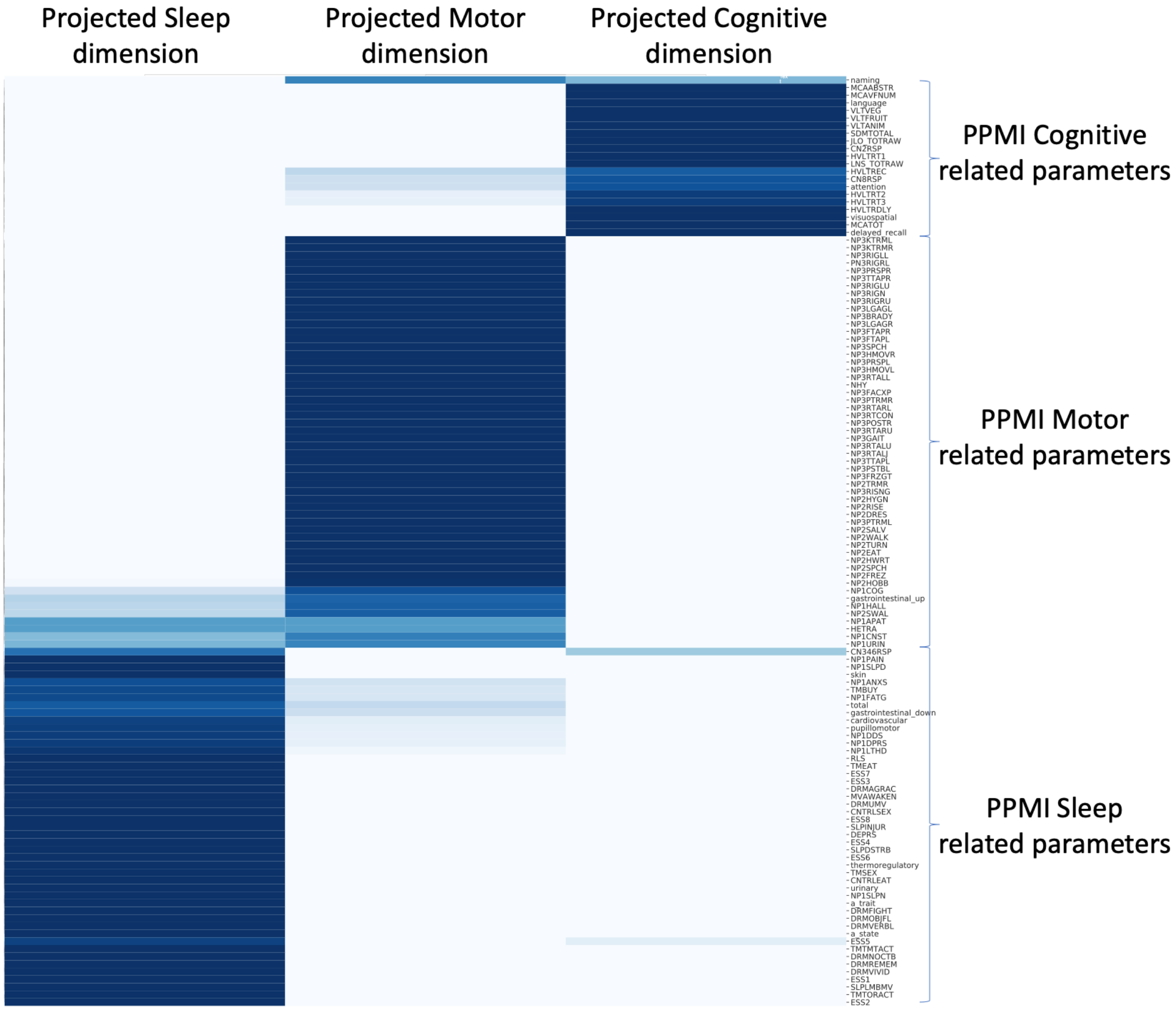


Figure 2.4: Shows how each 122 different input parameters have been projected into the new dimension of the Parkinson's progression space (cognitive, motor, and sleep dimensions). Darker colors represent strong mapping. The mapping is shown at the final visit after z-score normalization; it is similar for other visits as well (Figure not shown).

To understand the interpretation of PD progressions space dimensions, Figure 2.4 shows the mapping guide for how the PPMI's high-dimensional space of 122 different clinical parameters is mapped to the three-dimensional embedding of Parkinson's disease progression space. The features are grouped together to represent coherent skills. The leftmost component in Figure 2.4 mainly constitutes the questionnaire associated with sleep and mood problems, such as dream, fatigue, anxiety, and depression. The middle component represents questions related to motor skills such as speech, facial expression, tremor, and rigidity.

The third component represents questions related to cognitive skills, such as cognitive assessment and verbal learning tests. Therefore, the columns represent the projected three dimensions, i.e., motor, cognitive, and sleep-related trajectories, and the rows are the PPMI clinical parameters. This interpretable mapping is due to the property of NMF to group features showing similar variations in the data. This figure allows us to not only observe the conversion but also the heterogeneity of some clinical parameters, for instance how some of the Epworth Sleepiness Scale parameters reflect both sleep and cognitive disorders, and some reflect both sleep and movement disorders. We also looked at the features that seem to be incorrectly assigned, such as cognition (NP1COG) in the motor, and neurocranial (CN346RSP) in sleep. We find that the responses to these questions show minimal variation across subjects, which might make NMF assign them to any of the components. In comparisons of the eigenvalues within the NMF decomposition, the projected motor dimension was responsible for 63.58% of the explained variance, followed by the sleep dimension (21.81%), and cognitive dimension (14.61%).

2.3.3 Five year progression space

Figure 2.5 shows the disease trajectory of different PD subtypes. The progression space shows the gradual and linear change for all the subjects. Furthermore, the progression space tends to stabilize at the end of the third year. In this way, our model can capture patients' nuanced behavior showing their progression along with different skills. It is interesting to observe that a significant decline occurs between the second and third year for the subjects in our analysis. In terms of characteristics of PDs identified subtypes, Figure 2.6 demonstrated how cognitive, motor and sleep-related symptoms differ within each PDs subtype and in controls. There is a clear trend for increased cognitive, sleep, and motor disturbances after five years in fast progressors compared to the slower progressing subtypes. The slowest progressive subtypes (PDvec1) show a mild decline for motor dimension but less change for sleep and cognitive dimensions. We can observe that the difference in progression rates between controls and fastest progressive subtypes is mainly along the motor dimension followed by sleep and then the cognitive dimension.

Figure 2.7 shows the progression of each PD subtype overtime at baseline and after 12 months, 24, 36, 48 months, and 60 months. To better understand the clinical presentation of the three identified subtypes, Figure 2.7 and Figure 2.8 demonstrates the three main projected dimensions (motor, cognitive, and sleep-related disturbances), as well as actual clinical values of each subtype overtime for UPDRS-Part I, Part II, Part III, as well as Hopkins Verbal Learning Test, Symbol Digit Modalities Test, Semantic Fluency test, Epworth

Sleepiness Scale, State-Trait Anxiety Inventory for Adults, and Geriatric Depression Scale.

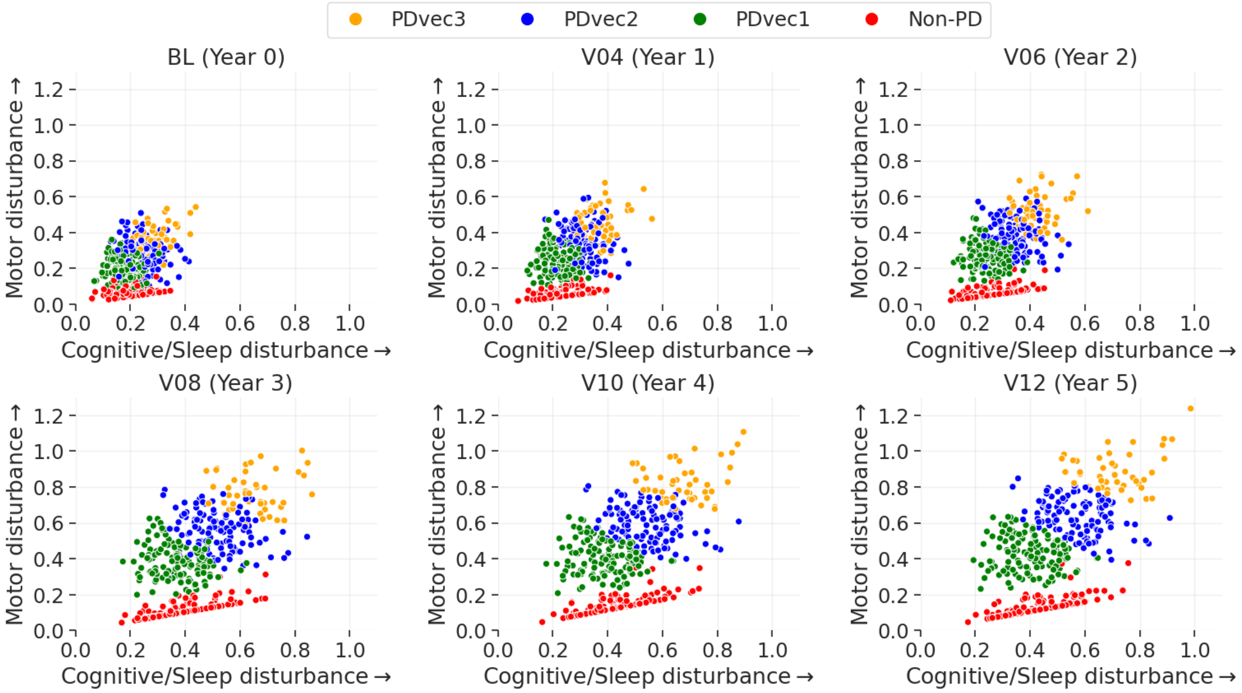


Figure 2.5: Visualization of two-dimensional progression space of PD subtypes at the end of every year, showing their normalized trajectory. (BL-baseline, V04-Year1, V06-Year2, V08-Year3, V10-Year4, V12-Year5).

2.3.4 Identified subtypes and their clinical characteristics

Figure 2.9 shows the visualization of unsupervised learning via Gaussian Mixture Model (GMM) in a two-dimension progression space. In two-dimensional progression space, the projected dimensions represent motor (y-axis) and cognitive (combined with sleep) (x-axis) components. Projected dimensions are normalized; the increase in values along either direction signifies a higher decline. GMM fits the data into different subtypes relating to velocity of decline across symptomatologies from non-PD controls. The Bayesian information criterion has identified three Gaussian distributions representing three PD subtypes. Further, mean values of PD subtypes in lower dimensional progression space are significantly different along both Motor dimension (PDvec1 = 0.43 [95%CI: 0.41-0.44], PDvec2 = 0.64 [95%CI: 0.62-0.65], PDvec3 = 0.89 [95%CI: 0.85-0.92)] and Cognitive/Sleep dimensions (PDvec1 = 0.40 [95%CI: 0.39-0.42], PDvec2 = 0.57 [95%CI: 0.56-0.59], PDvec3 = 0.71 [95%CI: 0.68-0.75]) all with non-overlapping CIs across groups. These three groups identified algorithmically within the case population change over time differently within the progression space

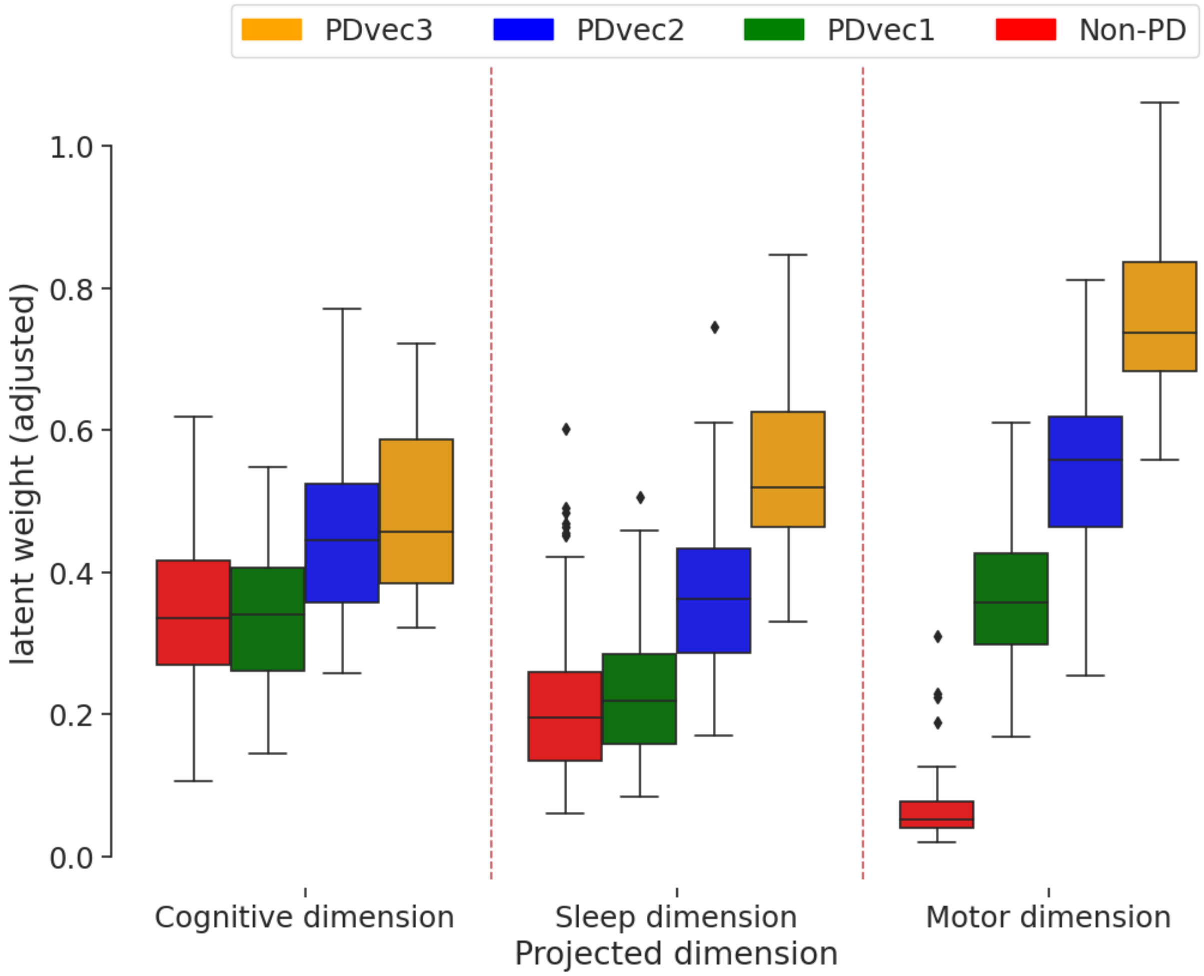


Figure 2.6: Shows the distribution of projected dimensions (cognitive, motor, and sleep) weights for each Parkinson's category and healthy control after five years. An increase in values along either direction reflects the increase in the disturbance. PDvec3 has the highest motor and sleep disturbance, as well as the highest cognitive impairment.

and across specific biomarkers of progression, with PDvec3 generally progressing at a much steeper slope (Figure 2.5, Figure 2.6, Figure 2.7, Figure 2.8). Using our proposed approach, 45% (134/294) of PD patients identified as PDvec1 (slow progressors), with 39% (114/294) belonging to PDvec2 (medium progressors) and PDvec3 (fast progressors) accounts for 16% (46/294) patients.

2.3.5 Biological characteristics of the identified subtypes

Figure 2.10 shows the variation of biological biomarkers for each PD subtype over time. In terms of patients' features, height and weight show a significant decline over time for the

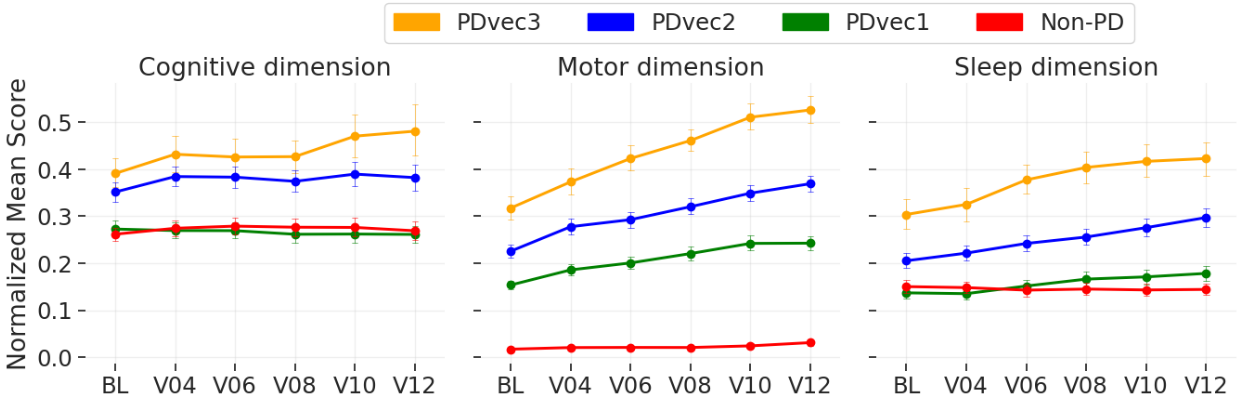


Figure 2.7: Shows the progression of each PD subtype over time for motor, sleep, and cognitive dimension overtime on the preprocessed values.

fast progressors (PDvec3) compared to other subtypes. We used a linear mixed effects model for association testing of PD subtypes and serum neurofilament light (Nfl) measurements. PDvec3 has a significantly steeper slope across time than PDvec1 after adjusting for sex, height, weight, and age at baseline ($P < 0.005$). Table 2.3 shows the association testing of Nfl with PD subtypes. PD patients have lower values compared to healthy controls for all CSF sample measurements such as alpha–synuclein, total tau protein, $A\beta\_42$, and p–Tau181.

2.3.6 Genetic analysis of the identified subtypes

In terms of the genetic association of PDs identified subtypes, genetic risk scores (GRS) were calculated [39, 53, 54]. As a one-time measurement, the GRS was not included during the longitudinal clustering exercise; however, we analyzed regressions comparing associations between the GRS and either the continuous predicted cluster membership probability (linear regression) or the binary membership in a particular cluster group compared to the others. All models were adjusted for age at onset, biological sex, and principal components as covariates to adjust for population substructure in PPMI. The GRS was significantly associated with decreasing magnitude of the sleep vector per Standard deviation (SD) of increase in the GRS (beta = -0.029, se = 0.010, p = 0.002, adjusted r2 = 0.046). For binary models of membership, we see that the GRS is weakly but significantly associated with a decreased risk of membership in PDvec3 (odds ratio = 0.563 per 1 SD increase from case GRS mean, beta = -0.574, se = 0.244, P = 0.018) and increased risk of membership in PDvec1 (odds ratio = 1.341, beta = 0.293, se = 0.134, P = 0.0282) all relative to the moderate progressing group as a reference. The lack of a strong genetic association is due to the small sample size, and that genetic variants relating to risk do not necessarily affect progression.

2.3.7 Replication in an independent cohort

In order to ensure the generalizability and validity of the results, we replicated the subtype identification in an independent PDBP cohort. Details on differences between the training (PPMI) and replication (PDBP) cohorts can be found in the Figure 2.2, Table 2.3. In the PDBP cohort, 46% (121/263) of PD patients were identified as PDvec1, 23% (60/263) belonged to PDvec2, and the remaining 31% (82/263) were classified into the PDvec3 group. We observed less manifested separation of PDvec2 (medium progressors) in the PDBP cohort. In the progression space, the spatial differences between subtypes become more apparent with increased longitudinal data. The reason can be attributed to the fact that the PDBP cohort has 3 years of longitudinal data compared to 5-year data in the PPMI cohort.

Figure 2.11 shows the identified subtypes in the independent PDBP cohort using the model developed on the PPMI dataset. We see that the identified subtypes in the PDBP cohort are similar to the ones in the PPMI dataset in terms of progression. Due to the limited length of the PDBP study (36 months), the visualization of progression space is shown through the 36 months follow-up from the baseline. The PPMI and PDPB cohorts are clinically different cohorts and recruited from different populations. The replication of our results in the PDBP cohort that was recruited with a different protocol shows the strength of our study's methodology. We demonstrate that if we ascertain the same phenotypes using standardized scales, we can reliably discern the same subtypes and progression rates. This suggests that our results may be generalizable and the clinical subtypes reproducible.

2.3.8 Supervised early subtype prediction

Following the data-driven organization of subjects into progression subtypes and clustering them into three subtypes, we developed three models to predict patient progression class after 60 months based on varying input factors: (a) from baseline clinical factors, (b) from baseline and year one clinical factor, (c) biological and genetics measurements. Figure 2.12a and Figure 2.12b show the ROC (Receiver Operating Characteristic) curves of our multi-class supervised learning predictors. We correctly distinguish patients with PD based on baseline only input factors and predict their 60-month prognosis with an average AUC of 0.92 (95% CI: 0.94 ± 0.01 for PDvec1, 0.86 ± 0.01 for PDvec2, and 0.95 ± 0.02 for PDvec3) at cross-validation. The predictor built on baseline and year 1 data performs even better with an average AUC of 0.953 (95% CI: 0.97 ± 0.01 for PDvec1, 0.91 ± 0.02 for PDvec2, and 0.97 ± 0.01 for PDvec3) also at cross-validation. In Figure 6d, we have shown the PD subtype predictive performance at baseline, only using baseline data, and years after, as more data becomes available and combined with the baseline. The increased accuracy trend is due to the availability of more information about a subject. This approach is also practical in a clinical setting, as physicians will provide a better prognosis for patients after a one-year follow-up. Out of identified top-20 features, 11 belong to the motor dimension, 5 are from the sleep dimension and 4 are a part of cognitive dimension, which is in line with the amount of variability explained by each dimension (Table 2.2). Further details on feature importance contributing to the accuracy of these models can be found in the section entitled Feature Importance and Figure 2.15, Figure 2.16 and Figure 2.17.

Table 2.2: Shows the top 20 clinical parameters used to obtain 0.92 AUC scores and their mapped dimension. Refer to Figure 2.15 for the scaled importance weights of each feature.

Feature	**Description**	**Latent dimension**
NHY	Hoehn and Yahr stage	Motor
NP3BRADY	Global Spontaneity of movement	Motor
NP3FACXP	Facial expression	Motor
NP2TRMR	Tremor	Motor
urinary	difficulty retaining urine + involuntary loss of urine + stream of urine been weak + pass urine at night + urine your bladder was not completely empty + urine again within 2 hours of the previous time	Sleep
SDMTOTAL	total symbol digit modalities test	Cognitive
VLTANIM	Total number of animals	Cognitive
NP1SLPD	Daytime sleepiness	Sleep
NP2HOBB	Doing hobbies and other activities	Motor
gastrointestinal_down	Have feeling during meal that you were full very quickly + Had problems with constipation + Had to strain hard to press stools + Had involuntary loss of stools	Sleep
NP3FTAPR	Finger tapping right hand	Motor
NP3RIGN	Rigidity – neck	Motor
HVTRT1	Immediate Recall Trial 1	Cognitive
NP2DRES	Dressing	Motor
NP3RTCON	Constancy of rest	Motor
NP2SALV	Saliva and drooling	Motor
DRMFIGHT	In my dreams: sudden limb movements	Sleep
DRMAGRAC	Dreams frequently have aggressive or action-packed content	Sleep
VLTVEG	Total number of vegetables	Cognitive
NP3PRSPR	Pronation-supination – right hand	Motor

Besides the cross-validation of predictive models in the PPMI cohort, we have also validated the accuracy of the predictive model in the independent PDBP cohort. The predictive model trained on the PPMI baseline data correctly distinguished PDBP patients with an

AUC of 0.84 (ROC curves in Figure 2.12c). The replicated predictive model performs very well for PDvec1 and PDvec3 (AUC of 0.91 and 0.88, respectively). However, due to the small sample size, the predictive model does not predict as well on PDvec2 (AUC of 0.73). Fewer samples make up the PDvec2 cluster in the replication cohort, and it has been easier for the predictive model to predict the more extreme subtypes (i.e., PDvec1 and PDvec3). Despite the smaller sample size of the PDBP cohort, the results strongly validate our previous observations of distinct, computationally discernible subtypes within the PD population. This finding indicates that our methodology is robust, and our unique progression analysis and clustering approach result in the same clusters. In summary, we have mined data to identify three clinically related constellations of symptoms naturally occurring within our longitudinal data that summarize PD progression (63.58%, 21.81%,14.61% variance loadings) comprised of factors relating to motor, sleep, and cognitive.

2.3.9 Biomarker based prediction

The performance of PD progression prediction models using biomarkers and genetic measurements for the PPMI cohort is shown in Figure 2.13 and Figure 2.14. A trained machine learning model using only UPDRS can predict subtypes with a 0.77 AUC score (Figure 2.13b) compared to 0.92 with the model that uses all symptomatic clinical measures at baseline on five-fold cross-validation. It demonstrates the utility of machine learning models in integrating features from multiple dimensions to provide an optimal classification performance. Biomarkers, such as age, height, weight, and CSF measurements, are shown to be essential features in predicting the subtypes at baseline (shown in Figure 8). The mean AUC score is 0.67 (Figure 2.13e) using biospecimen, vital signs, and demographics. In comparison, genetic features show slightly lower performance (AUC score 0.66, Figure 2.13d). A combination of demographics, biospecimen, vital signs, genetics, and UPDRS is the best performing model (AUC score 0.80, Figure 2.13a). It is important to note that segregating PDvec3 (fast progressive subtype) has shown similar performance with only the UPDRS model and that model that includes other biomarkers and genetic measurement. It might be valuable to evaluate the UPDRS model's performance in a clinical setting, as UPDRS is a standard measure of PD diagnosis and disease severity. Further, the individual components (clinical questionnaire responses) of UPDRS are crucial, and machine learning models exploit and utilize their complex interaction to form a composite score of PD subtypes prediction. A simple aggregation (average) of UPDRS individual responses might not have similar subtype prediction power. Based on the ease of availability in real-world clinical settings, we suggest combining UPDRS, genetics, biomarkers, and demographics as a subtype diagnostic model, which has a 0.80 AUC score (Figure 2.13a). Further validation of the model is necessary to improve generalizability in other cohorts to make it an application for clinicians.

2.3.10 Feature importance

The predictive model was also analyzed to identify the feature importance in predicting PD subtypes. Feature importance is determined by calculating the relative influence of each variable, which is typically given by information gain/entropy, and how much the variable contributes to the accuracy (Figure 2.15). We further scaled each feature's importance between 0 and 1 using min-max normalization. Figure 2.15 shows the top-50 features identified by our predictive model. We list the top 20 features used as input to obtain 0.92 AUC with an ensemble of machine learning models.

SHAP is an unified approach to explain the output of any supervised machine learning model. It assigns an importance value to every feature based on Shapley values. In addition, it generates the impact of each feature on the model's output i.e. the class probability for classification algorithms. The best performing model among five folds is chosen to calculate the SHAP values. Figure 2.16 (left) shows the impact score (SHAP contribution) on the probability of PDvec3 class. We see that a higher serum_nfl score corresponds to the increase in the probability of a patient belonging to the PDvec3 class. Similarly, higher scores on other symptomatic features related to hobby, sleep are among the top features that can differentiate between PD_h and other classes. Figure 2.16 (right) shows the behavior of the model for PDvec1 class. The top features include sleeping behavior, Hoehn and Yahr stage score and the posture stability with probability of lower progressive class increases with increase in scores for these features. Younger PD patients at screening are expected to show lower PD progression as compared to the older patients. Figure 2.17 shows the top 20 features involved in classifying PD progressive subtypes.

2.3.11 Change in diagnosis status

The clinical condition of patients can deteriorate, stay the same, or rarely gets better with time. As the study progresses and more information becomes available about the disease manifestation, the patient's diagnosis will be updated. We looked at cases where their clinical diagnosis were updated in the PPMI study. Figure 2.18, shows the trajectory of two patients initially diagnosed as PD in the progression space. We can observe that the patient whose status has changed from PD to dementia has much worse condition along the Cognitive dimension. The other patient whose status changed from PD to multiple system atrophy has shown more decline along the motor dimension.

2.3.12 Association testing of Nfl with PD subtypes

Table 2.3: Shows the longitudinal changes in serum Nfl levels over 5 years for three subtypes. We used a statistical t-test between PDvec1 vs. PDvec2 and PDvec1 vs. PDvec3 to compare the means of slope and serum Nfl levels at different points in time.

Outcome	PDvec1	PDvec2	PDvec3
Change in serum Nfl level per year Mean [SD]	1.31 [2.36]	1.48 [2.76]	2.91 [4.23]
t-test **P-value [t-statistic]**	-	0.6196 [-0.50]	0.0025b [-3.07]
Baseline Mean [SD]	11.54 [5.84]	11.77 [5.21]	15.42 [6.66]
t-test **P-value [t-statistic]**	-	0.7576 [-0.31]	0.0004 [-3.63]
At end of year1 Mean [SD]	11.72 [5.05]	13.69 [10.42]	15.72 [6.38]
t-test **P-value [t-statistic]**	-	0.086 [-1.73]	0.0002 [-3.86]
At end of year2 Mean [SD]	13.34 [8.09]	13.85 [7.17]	17.09 [7.91]
t-test **P-value [t-statistic]**	-	0.6321 [-0.48]	0.0138 [-2.49]
At end of year3 Mean [SD]	14.16 [8.98]	14.76 [8.98]	19.87 [8.99]
t-test **P-value [t-statistic]**	-	0.5761 [-0.56]	0.0006 [-3.51]
At end of year5 Mean [SD]	16.72 [13.51]	18.02 [12.36]	26.75 [20.41]
t-test **P-value [t-statistic]**	-	0.4435 [-0.77]	0.0003 [-3.73]

Table 2.3 below details baseline and follow-up differences in Nfl across the predicted progression vector classes. Here we show significant differences between slow and faster progressors not only in the measures of Nfl itself but in the slope of change. Correcting for relevant parameters using a linear mixed-effects model In addition to subtype, Nfl measurements might be sensitive to other factors such as sex, height, weight, and age at baseline. We used a linear mixed effects model to test for the association between subtypes and Nfl measures after adjustment.

2.4 DISCUSSION

Prediction of disease and disease course is a critical challenge in the patient counseling, care, treatment, and research of complex heterogeneous diseases. Within PD, meeting this challenge would allow appropriate planning for patients and symptom-specific care (for example mitigating the chance of falls, identifying patients at high risk for cognitive decline or rapid progression, etc.). Perhaps even more importantly at this time, prediction tools would facilitate more efficient execution of clinical trials. With models predicting a patient-specific disease course, clinical trials could be shorter, smaller, and would be more likely to detect smaller effects, thus, decreasing the cost of phase 3 trials dramatically and essentially reducing the exposure of pharmaceutical companies to a typically expensive and failure-prone area. We previously had considerable success in constructing, validating, and replicating a model that allows a data-driven diagnosis of PD and the differentiation of PD-mimic disorders, such as those patients who have parkinsonism without evidence of dopaminergic dysfunction [59]. We set out to expand this work by attempting to use a novel approach to 1) define natural subtypes of the disease, 2) attempt to predict these subtypes at baseline, and 3) identify progression rates within each subtype and project progression velocity.

While the work here represents a step forward in our efforts to sub-categorize and predict PD, much more needs to be done. The application of data-driven efforts to complex problems such as this is encouraging; however, the primary limitation of such approaches is that they require large datasets to facilitate model construction, validation, and replication. These datasets should include standardized phenotype collection and recording to achieve the most powerful predictions. Longer follow-ups, more ancestral diversity in samples, and large sample series are crucial to broadening the applicability of this work. Collecting such data is a challenge in PD, with relatively few cohorts available with deep, wide, well-curated data. Thus, a critical need is the expansion or replication of efforts such as PPMI or PDBP, importantly with a model that allows unfettered access to the associated data; the cost associated with this type of data collection is large, but these are an essential resource in our efforts in PD research. Global Parkinson's Genetics Program (GP2) project has the potential to address some of these limitations in the future (https://gp2.org/).

A study used cluster analysis to identify patient subtypes and their corresponding progression rates [23], although these used percentile cutoffs and are not completely data-driven in nature. However, this study evaluated clusters according to only two-time points, baseline, and short-term follow-up, that were aggregated into a Global Composite Outcome score. In return, the subtypes did not capture the fluctuations in the prognosis of subtypes. More recently, a study used a Long-Short Term Memory-based deep learning algorithm to discover PD subtypes, with each subtype showing different progression rate18. The loss of interpretability with deep learning models makes their approach less suitable for practical purposes. Another study proposed a trajectory-based clustering algorithm to create patient clusters based on trajectory similarity [81]. The algorithm gives equal importance to all the features; however, PD is a multi-dimensional spectrum of symptoms with overlapping features derived from simultaneous pathological processes [82]. Finally, in order to be used in practice, subtyping solutions need to be replicated in a different cohort to show the reliability of methods in assigning individual patients to a subtype. Additionally, none of these previous studies used completely independent replication data.

Our findings can also have implications for the day-to-day practice of clinicians. Movement disorders specialists often use screening tools such as MDS-UPDRS to assess a patient's progression and response to treatment. However, performing these clinical assessments requires experience, expertise, and time, which hinders its widespread use by other clinicians (and even neurologists who are not trained in movement disorders). Underutilization of clinical assessment tools can lead to the suboptimal characterization of PD patients and their clinical course, which in turn impacts their care. Our study is one of the first of its kind which systematically assessed the accuracy of each feature of MDS-UPDRS in predicting PD's course. For example, daytime sleepiness (NP1SLPD) was found to have the highest importance in clinical progression, followed by doing hobbies and activities (NP2HOBB), dressing (NP2DRES), and urinary problems (NP1URIN). Knowing the clinical features with the highest yield in course prediction can help clinicians to tailor their assessment and better inform patients about their disease course. In addition, shortened versions of comprehensive assessment tools can be utilized to address specific clinical questions. Surprisingly, none of the genetic markers studied had high accuracy in clinical course prediction. Our work initiates multiple questions that are worth exploring in the future. The progression space seems to stabilize after three years from the baseline. It will be interesting to predict how much time (from baseline) is required to provide reliable predictions about the PD subtypes. Secondly, fast progressors do not worsen with the multiple symptoms such as Epworth and MDS-UPDRS scores (Figure 2.8) from the fourth to the fifth year, while other subtypes do. It raises the question of whether the fast progressors reach the saturation point after some time from baseline. It will be useful if we can look for similar patterns in other PD datasets. Incorporating imaging data for PD subtypes is also an exciting direction to pursue in the future. Finally, dramatic increases in Nfl and high baseline levels of Nfl could be an indicator of potential rapid progression.

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

This work provides data-driven subtypes in distinct progression stages of PD and discusses an approach to predict the future rate of progression years from baseline using longitudinal clinical data. Predicting disease progression is a paramount challenge in treating and curing several complex diseases. This study is a step forward toward designing sophisticated machine-learning paradigms to facilitate the early diagnosis of PD progression and longitudinal biomarker discovery such as our finding of elevated Nfl in fast progressors (both at baseline and with regard to the rate of change per year). Predicting PD progression rates would lead to better patient-specific attention by recognizing the patients with a swift rate of progression at an early stage. The proposed disease progression and trajectory prediction algorithms can help healthcare providers to develop a methodical and organized course for clinical tests, which can be much more concise and effective in detection. These adaptations and modifications in clinics may help to diminish treatment and therapy costs for PD. Further, the capability to anticipate the trajectory of impending PD progression at the early stages of the disease is an advancement toward uncovering novel treatments for PD modification. The proposed analysis provides insights to inhibit or decelerate the progression of PD-related symptoms and subsequent deterioration in the characteristics of life that are accompanied by the disease.

We have demonstrated that applying machine learning provides a systematic, data-driven way to understand disease heterogeneity. We discussed how unsupervised learning and visualization in lower-dimensional space can provide insights into disease understanding. However, these approaches have some limitations. Firstly, the linear nature of non-negative matrix factorization (NMF) cannot capture nuances in the data. Secondly, in this work, the dimensionality of the data in this chapter is in the order of 100s, but other multimodal biological datasets can consist of more than 1000 features. In the next chapter, we explore the applicability of a non-linear, graph-based dimensionality reduction tool called Aligned UMAP on large longitudinal biomedical datasets.

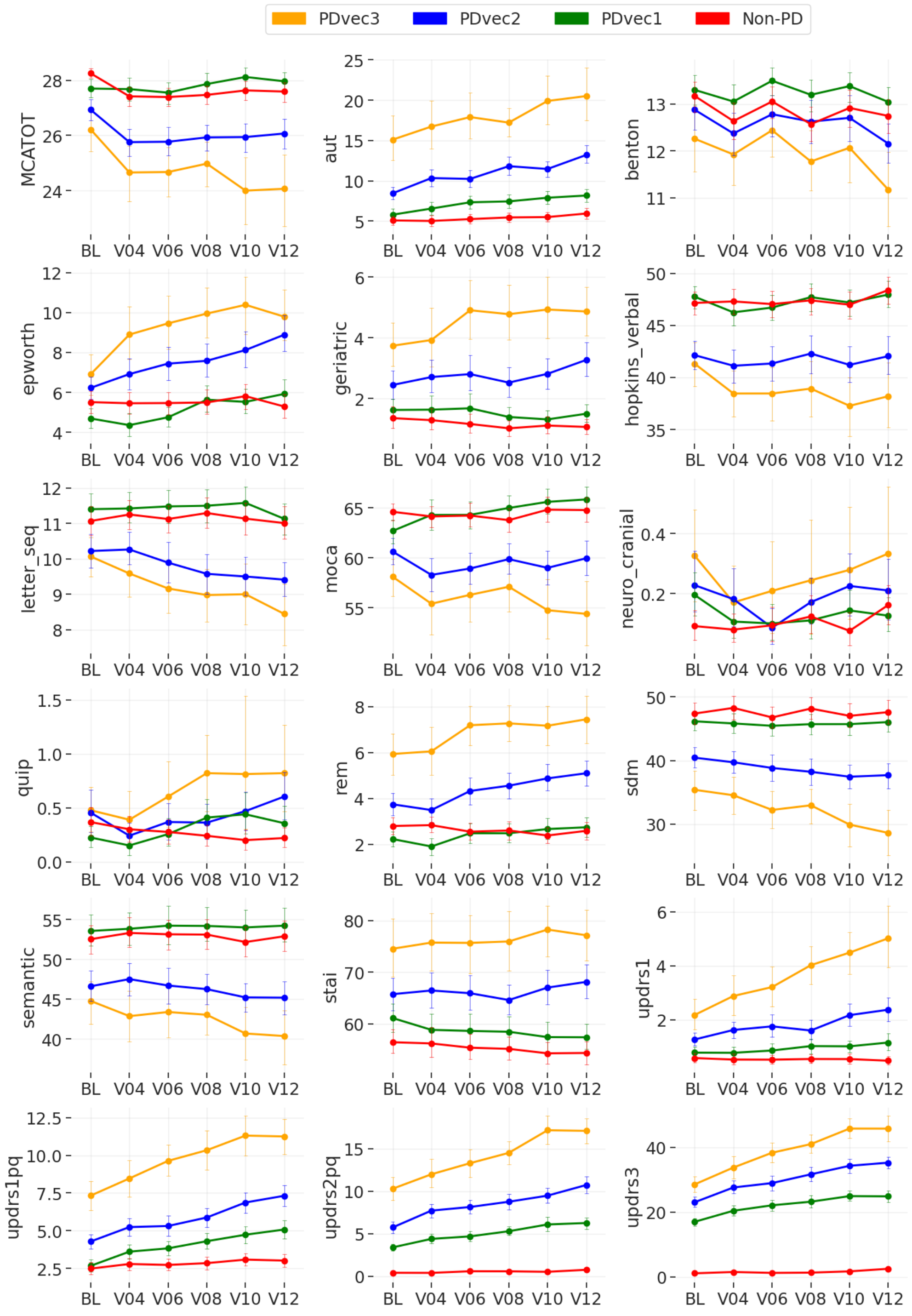


Figure 2.8: Shows the progression of each PD subtype over time. The graphs demonstrate the actual clinical values of each subtype overtime for UPDRS-Part I, Part 2, Part 3, as well as Hopkins Verbal Learning Test, Symbol Digit Modalities Test, Semantic Fluency test, Epworth Sleepiness Scale, State-Trait Anxiety Inventory for Adults, and Geriatric Depression Scale. BL: Baseline. V04: visit number 4 after 12 months. V06: visit number 6 after 24 months. V08: visit number 8 after 36 months. V10: visit number 10 after 48 months. V12: visit number 12 after 60 months.

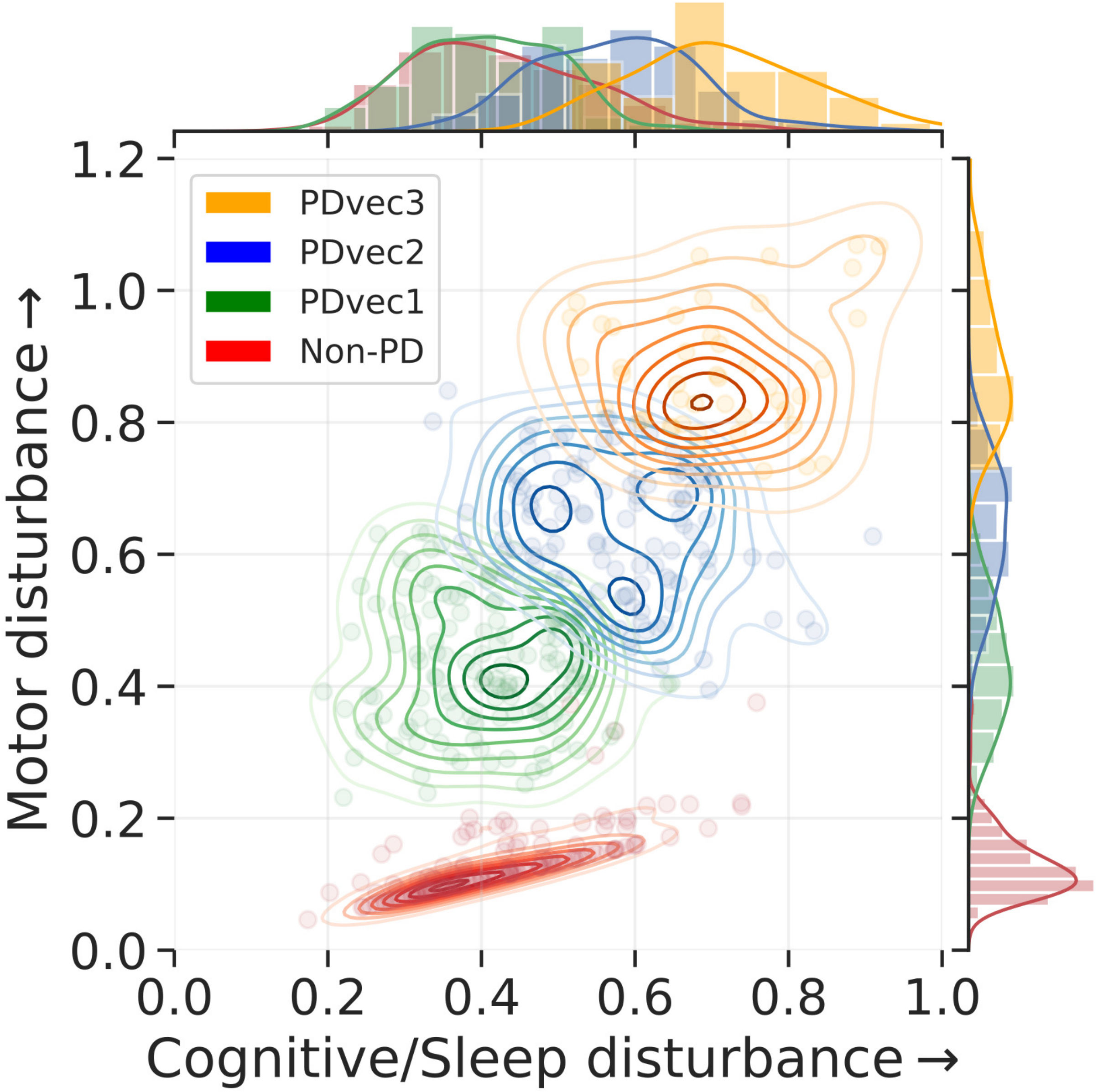


Figure 2.9: **PD five-year progression space**. Visualization of unsupervised learning via GMM on two-dimensional progression space and identification of three Gaussian distributions representing three distinct PD subtypes. An increase in value along either direction reflects the increase in the disturbance on a normalized scale.

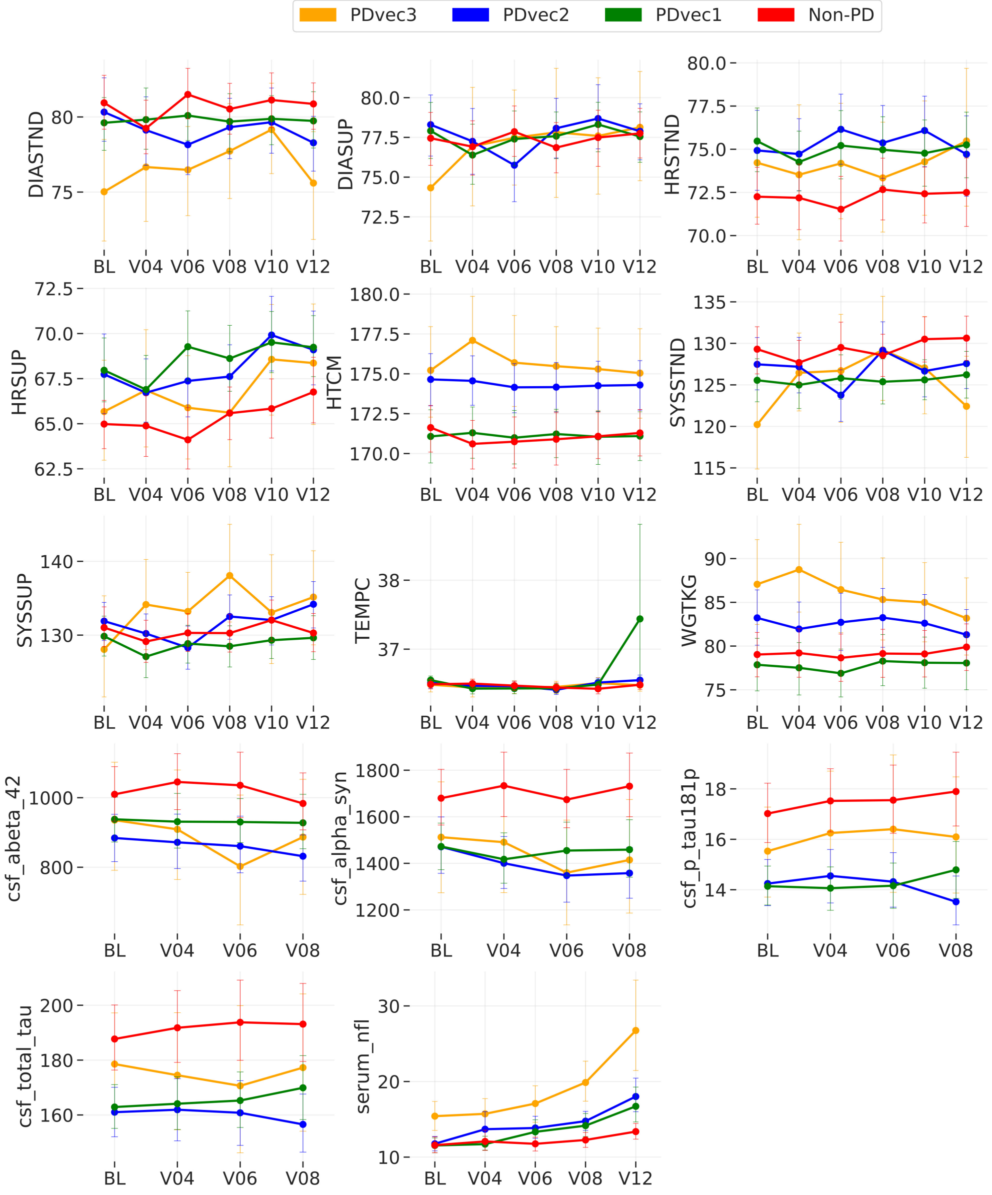


Figure 2.10: **Shows the biological biomarker variation of each PD subtype over time**. The graphs demonstrate the actual clinical values of each subtype overtime for vital signs (DIASTND standing diastolic blood pressure (BP), DIASUP supine diastolic BP, HRSTND standing heart rate, HRSUP supine heart rate, SYSSTND standing systolic BP, SYSSUP supine systolic BP, HTCM height in cm, TEMPC: temperature in C, WGTKG weight in kg), cerebrospinal fluid (abeta_42 beta-amyloid 1–42, alpha_syn alpha-synuclein, p_tau181p phospho-tau181, total_tau total tau protein), and serum neurofilament light levels (serum_nfl). BL: Baseline. V04 visit number 4 after 12 months. V06: visit number 6 after 24 months. V08 visit number 8 after 36 months. V10 visit number 10 after 48 months. V12 visit number 12 after 60 months. In all panels, data is presented as mean.

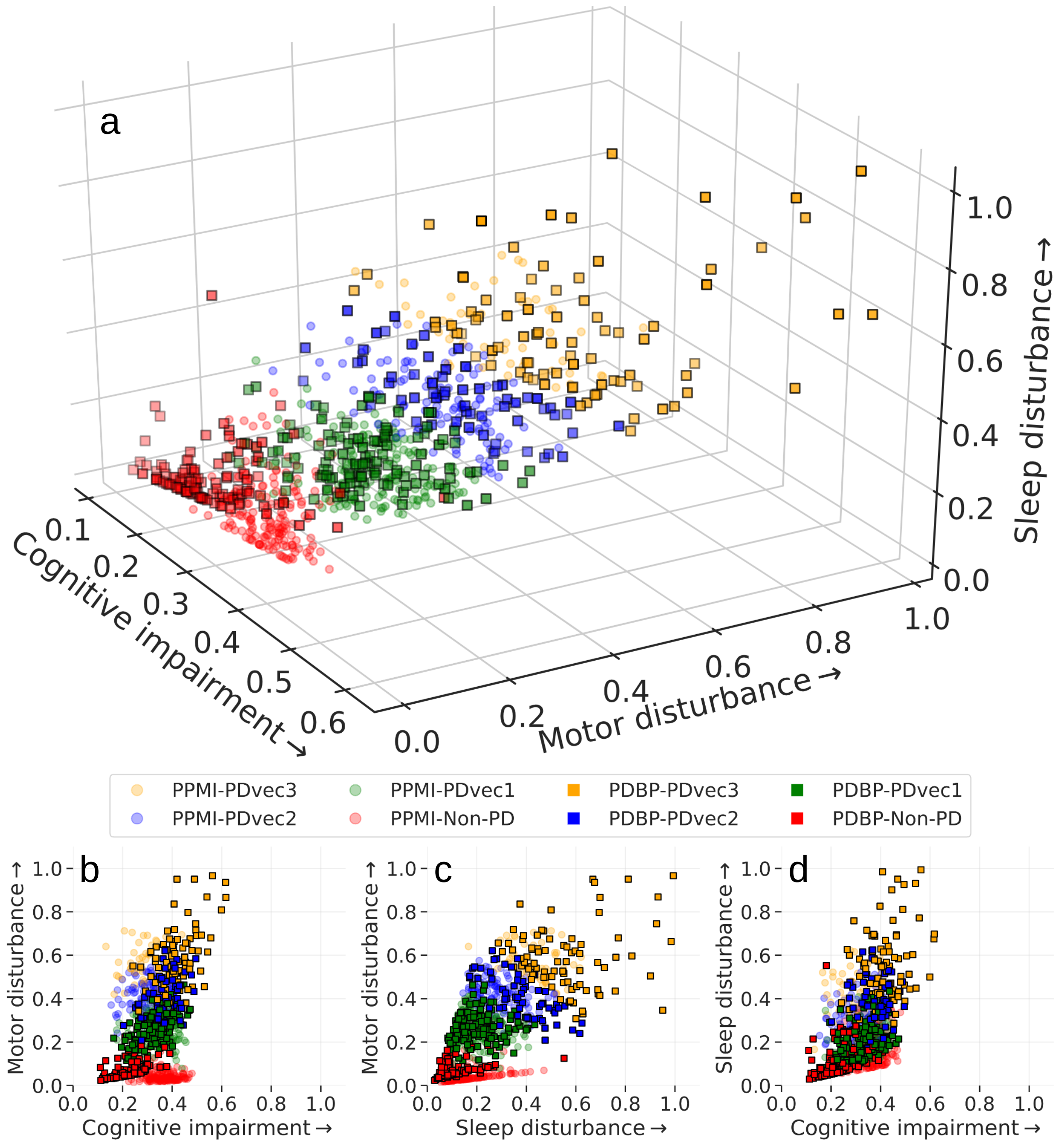


Figure 2.11: Shows the identified subtypes in the independent PDBP cohort using the model developed on the PPMI dataset. Similar PDBP and PPMI subtypes in terms of progression. a Shows the view of all three dimensions, b view of the motor and cognitive dimensions, c view of motor and sleep dimensions, and d view of sleep and cognitive dimensions. The normalized progression space is shown through the 36 months follow up from baseline for both PPMI and PDBP datasets.

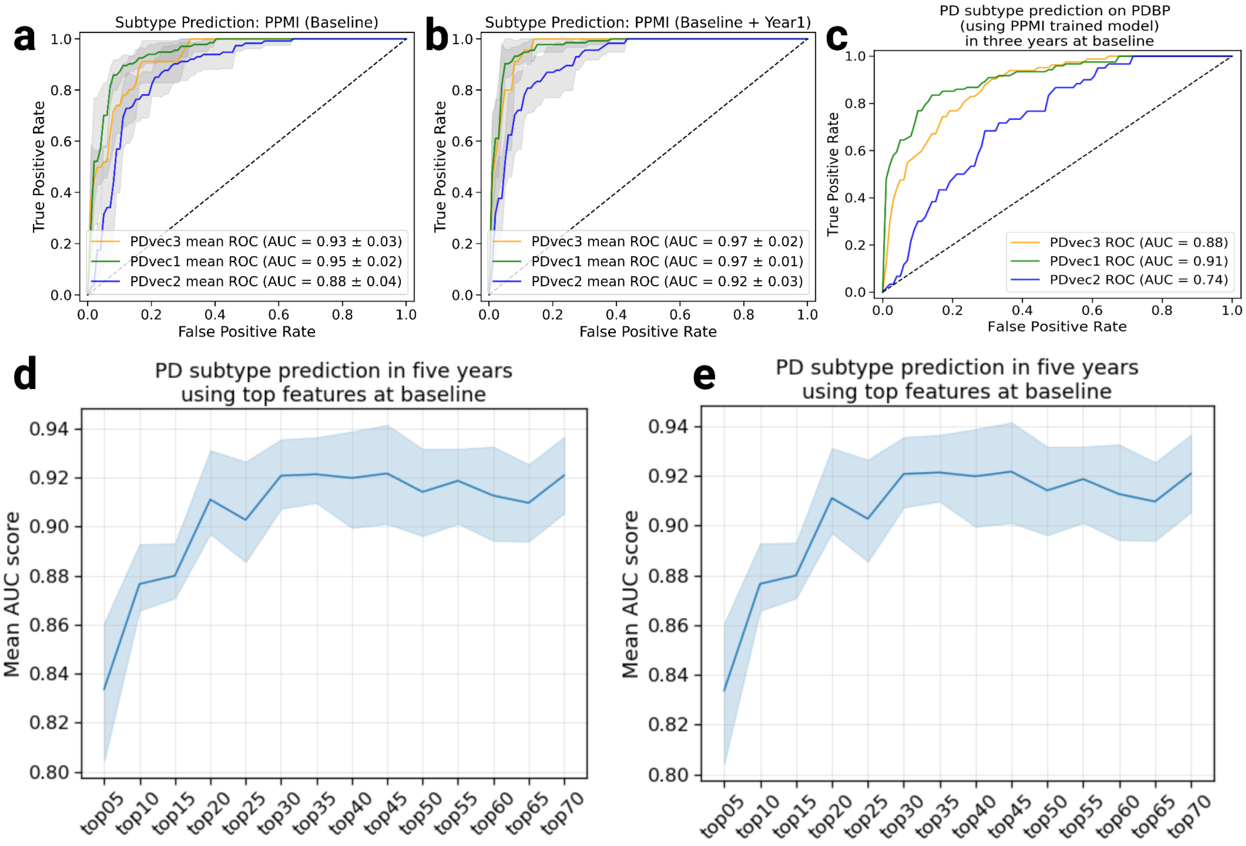


Figure 2.12: Shows the performance of Parkinson's disease progression prediction models. a The ROC (receiver operating characteristic) for the predictive model at baseline developed on the PPMI cohort evaluated using five-fold cross-validation. b The ROC for the predictive model developed on the baseline, and first-year data of the PPMI cohort evaluated using five-fold cross-validation. c The ROC for the predictive model developed on the PPMI baseline and tested on the PDBP cohort. d Performance of predictive models using data starting from baseline, only using baseline data, and years after, as more data becomes available and combined with the baseline. The y-axis shows the average AUC score across PD subtypes in the PPMI dataset. e Contribution of important features to achieve high accuracy. By including only 20 features, we can achieve an AUC of greater than 0.90. In all panels, data is presented as mean.

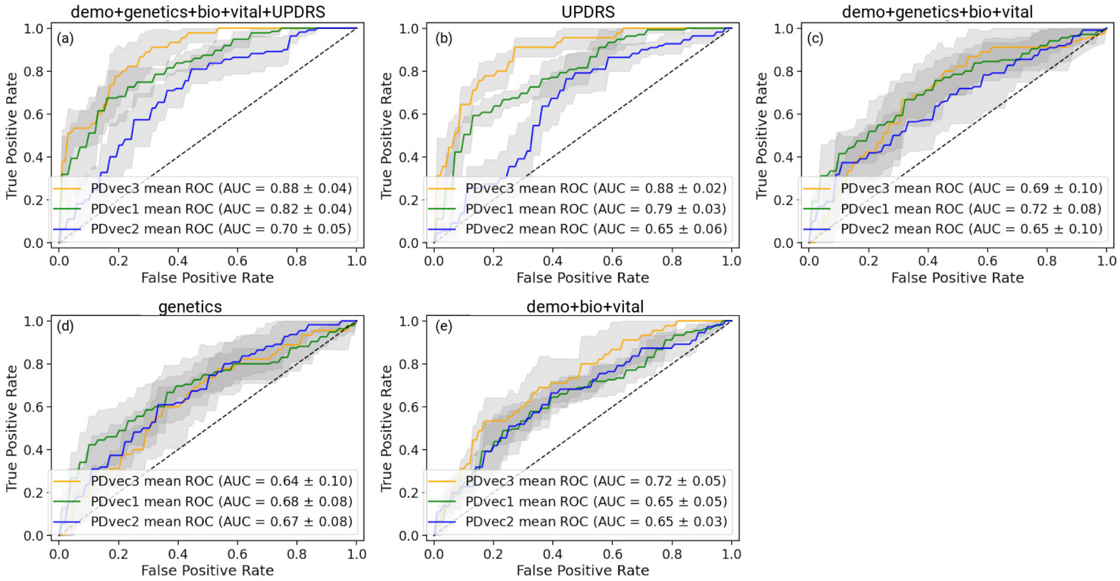


Figure 2.13: Shows the performance of Parkinson's disease progression prediction models using biomarkers and genetic measurements for the PPMI cohort. All models are evaluated using five-fold cross-validation. From top left to bottom right: a The ROC for the predictive model using a combination of demographics (education, year, sex, race), biospecimen (cerebrospinal fluid, serum Nfl levels), genetics (hg genotype), vital signs (weight, height, blood pressure) and UPDRS measurements. b The ROC for the predictive model developed on UPDRS scores. c The ROC for the predictive model developed using demographics, genetics, vital signs, and biospecimen measurements. d The ROC for the predictive model developed on genetic measurements e The ROC for the predictive model uses only demographics, vital signs, and biospecimen measurements. In all panels, data is presented as mean.

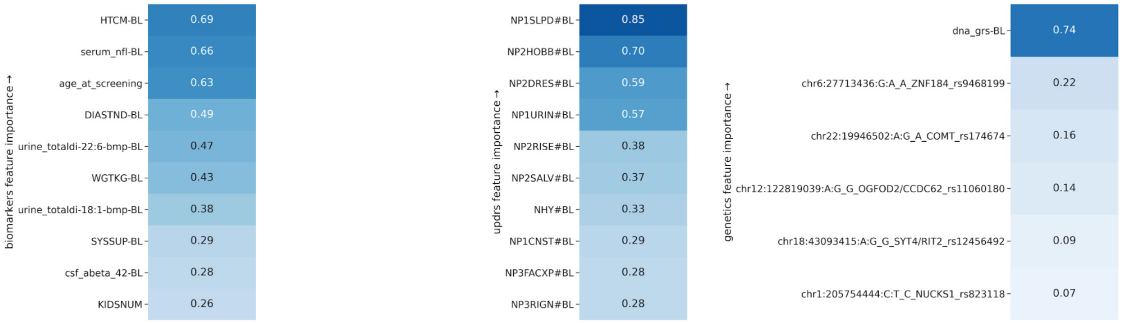


Figure 2.14: Heatmap plot showing significant contributing clinical parameters based on demographics, vital signs, baseline biospecimen, baseline MDS-UPDRS scores, and genetic measurements. The importance score of each feature is relative. BL baseline, HTCM height in cm, serum_nfl serum neurofilament light levels, age_at_screeing Age at screening, DIASTND standing diastolic blood pressure (BP), urine_totaldi urine levels of di-22:6-bis (monoacylglycerol) phosphate, WGTKG weight in kg, SYSSUP supine systolic BP, csf_abeta_42 cerebrospinal fluid β-amyloid 1–42, KIDSNUM number of kids, dna_grs genetic risk score.

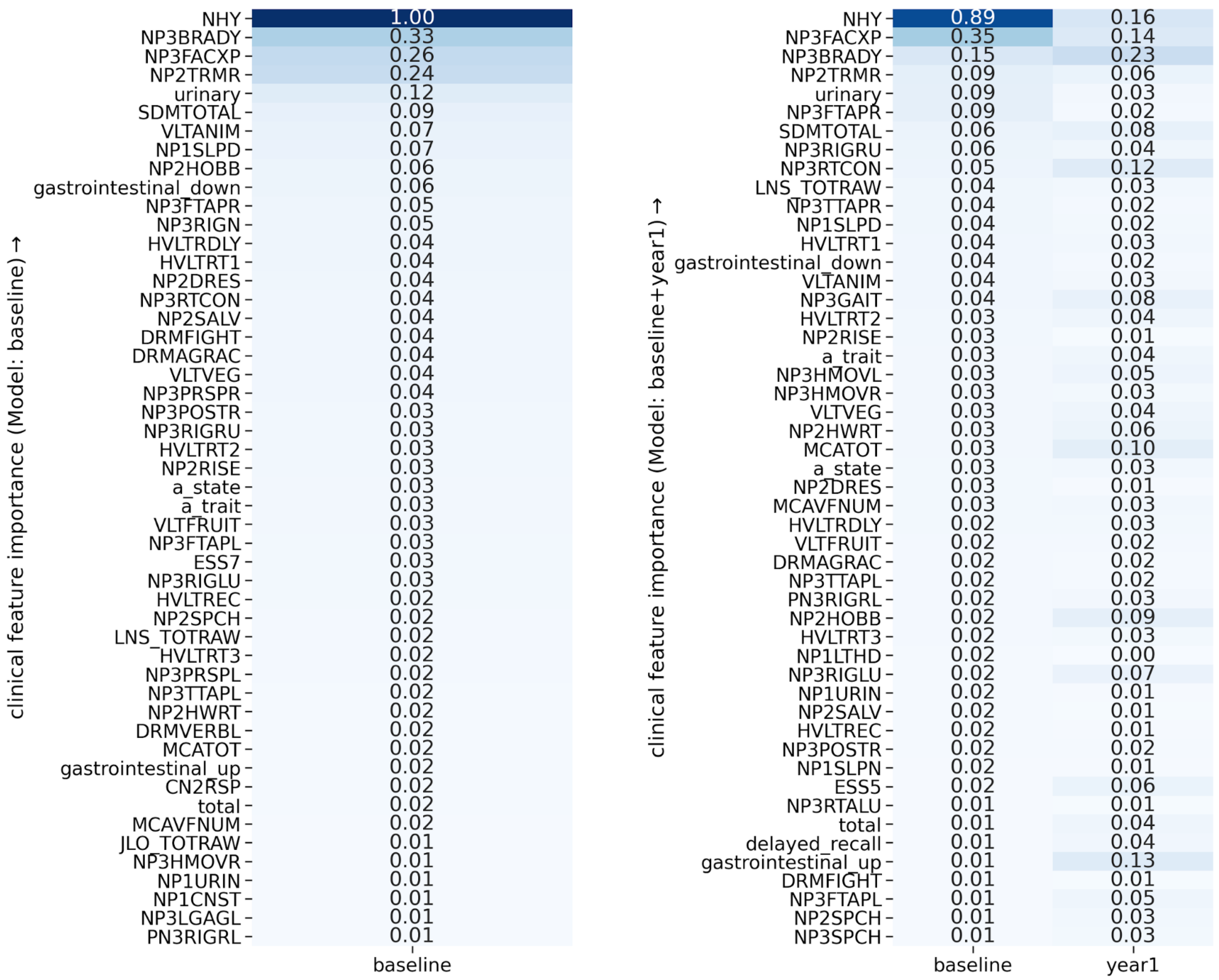


Figure 2.15: Shows the summary of clinical parameters (top 50 features) to the prediction models ordered by their importance. The value indicates the scaled importance of the variables in predicting the PD subtypes. Table lists significantly contributing clinical parameters based on the baseline model and on model using both baseline and year 1 data.

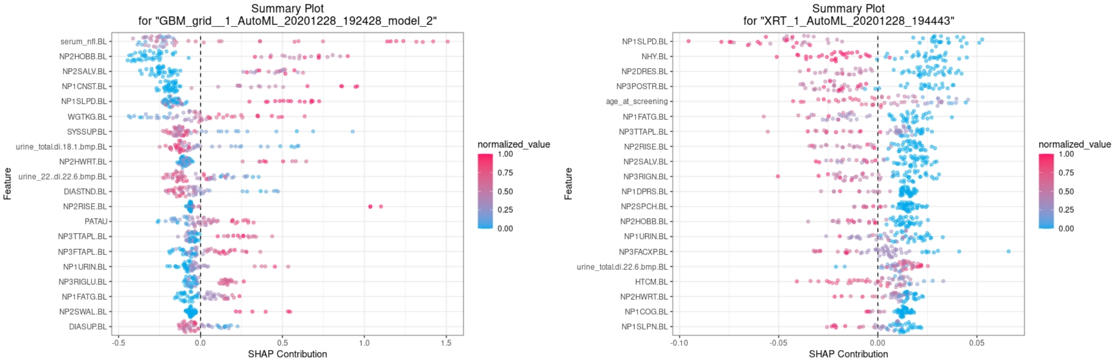


Figure 2.16: Clinical features influencing Parkinson's Disease progression class. Panels A and B from left to right. Detailed view of influence of top features for Higher PD progression class i.e. PDvec3 (Left) and lower PD progression class i.e. PDvec1 (Right). Higher value on the horizontal axis represents higher probability of a PD patient belonging to the PDvec3 class (left) and PDvec1 class (right).

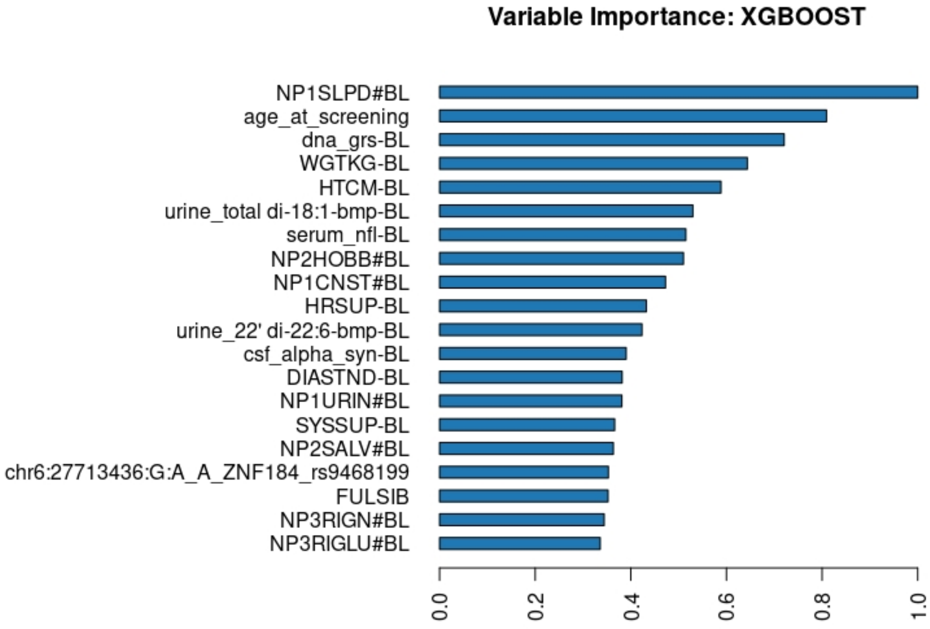


Figure 2.17: Clinical features influencing Parkinson's Disease progression class. Panels A and B from left to right. Detailed view of influence of top features for Higher PD progression class i.e. PDvec3 (Left) and lower PD progression class i.e. PDvec1 (Right). Higher value on the horizontal axis represents higher probability of a PD patient belonging to the PDvec3 class (left) and PDvec1 class (right).

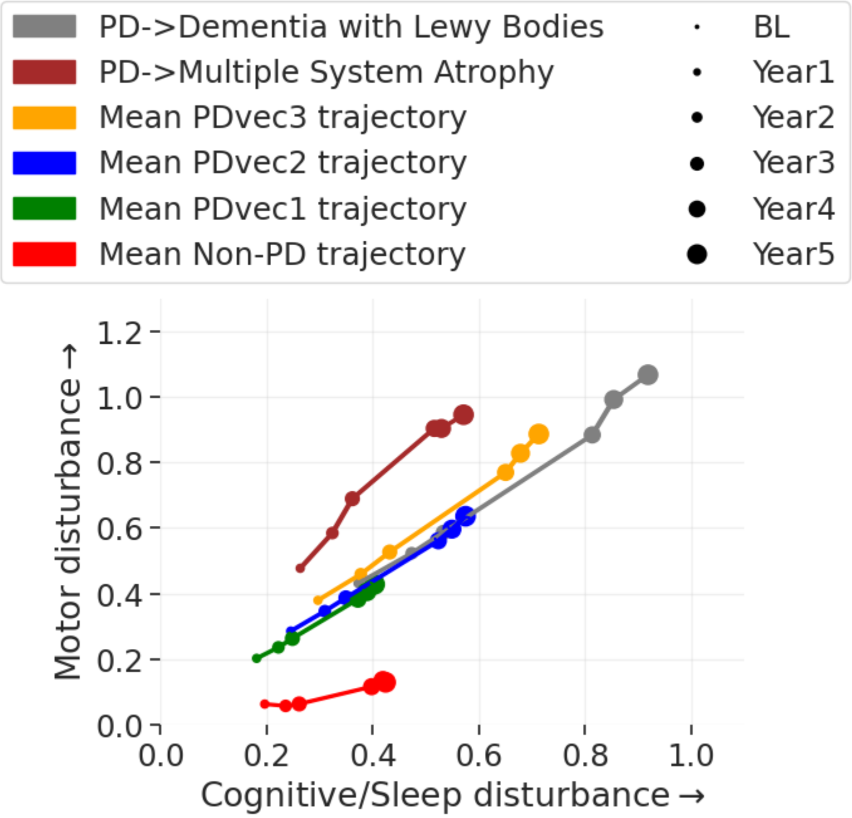


Figure 2.18: Shows the trajectory of two PD patients in the two dimensional progression space whose status has changed from their recruitment category in PPMI cohort. It also shows the average trajectory of PD subtypes and Non-PD subjects. The marker size corresponds to time from baseline.

CHAPTER 3: APPLICATION OF ALIGNED-UMAP TO LONGITUDINAL BIOMEDICAL STUDIES

In the previous chapter, we explored a use case of unsupervised learning algorithms and demonstrated how visualizing data in lower dimensions can help us understand disease. In this chapter, we aim to extend this approach to other longitudinal biological datasets and explore new use cases beyond disease heterogeneity.

3.1 INTRODUCTION

High-dimensional data analysis starts with projecting the data to low dimensions to visualize and understand the underlying data structure. Several methods have been developed for dimensionality reduction, but they are limited to cross-sectional datasets. Recently proposed Aligned-UMAP, an extension of the UMAP algorithm, can visualize high-dimensional longitudinal datasets. We demonstrated its utility for researchers to identify exciting patterns and trajectories within enormous data sets in biological sciences. We found that the algorithm parameters also play a crucial role and must be tuned carefully to utilize the algorithm's potential fully. We also discussed key points to remember and directions for future extensions of Aligned-UMAP. Further, we made our code open-source to enhance the reproducibility and applicability of our work. We believe our benchmark study becomes more important as more and more high-dimensional longitudinal data in biomedical research becomes available.

3.1.1 Bigger picture

Longitudinal multi-dimensional biological datasets are ubiquitous and highly abundant. These datasets are essential to understanding disease progression, identifying subtypes, and drug discovery. Discovering meaningful patterns or disease pathophysiologies in these datasets is challenging due to their high dimensionality, making it difficult to visualize hidden patterns. In this work, we applied Aligned-UMAP on a broad spectrum of clinical, imaging, proteomics, and single-cell datasets. Aligned-UMAP reveals time-dependent hidden patterns when color-coded with the metadata. Altogether, based on its ease of use and our evaluation of its performance on different modalities, we anticipate that Aligned-UMAP will be a valuable tool for the biomedical community.

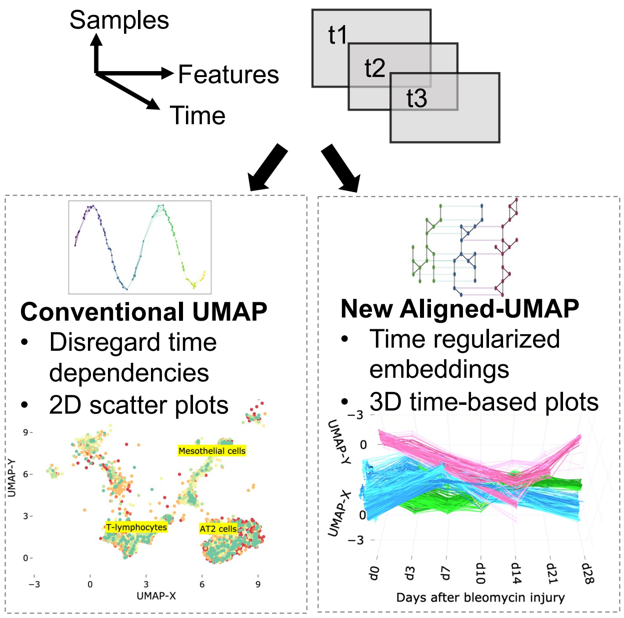


Figure 3.1: Graphical Abstract.

3.1.2 Highlights

The key highlights of our work are as follows:

- explored the utility of Aligned-UMAP in longitudinal biomedical datasets
- offer insights on optimal uses for the technique
- provide recommendations for best practices

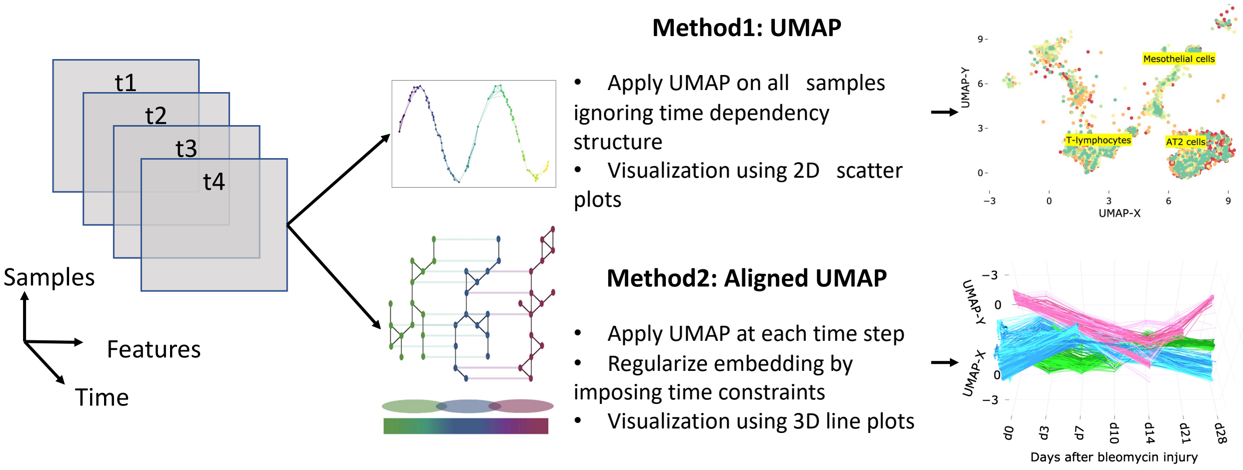
Visualizing large-scale, high-dimensional datasets is the starting step for any data exploratory analysis. Visualizing data is particularly useful for the biological community, where researchers rely on hypothesis-free data-driven analytics to gain essential insights and observe meaningful patterns from the data. The standard way of visualizing high-dimensional data is to project the data into low-dimensional space, typically 2D or 3D, while preserving local and global relationships. This transformation is called dimension reduction and belongs to the unsupervised machine learning algorithms class. The lower-dimensional data space can guide us in various tasks, such as identifying clusters, sub-structures, and outliers, detecting batch effects, and quality control measures to perform reliable and accurate downstream analyses.

In contrast to traditional methods for dimensionality reduction—for example, principal component analysis (PCA) [83] — Uniform Manifold Approximation and Projection

(UMAP) [84] learns a nonlinear embedding of the original space by optimizing the embedding coordinates of individual data points using iterative algorithms. It aims to accurately preserve the original local neighborhood of each data point in the visualization. Because of the expressiveness of nonlinear embeddings, UMAP is well regarded for its state-of-the-art empirical performance at elucidating sophisticated manifold structures. The biomedical community widely adopts UMAP for multiple studies ranging from Single Cell RNAseq data [85] to Genetics [52, 86] or complex clinical symptoms [32, 85] to depict exciting patterns from the data. In these use cases, UMAP is explored on datasets assuming that all samples in the dataset are independent.

Despite the prevalence of non-independent high-dimensional biological datasets, the application of UMAP in this area is little explored. This non-independence effect can occur from measurements at different time intervals, age, or other discrete/continuous variables. There are various longitudinal datasets of different modalities such as Clinical Symptoms, Magnetic Resonance Imaging (MRI), Electronic Health Records (EHR), Electroencephalography (EEG) for sleep monitoring, Electrocardiogram (ECG) data, etc. Since UMAP is a stochastic algorithm, different runs with the same hyperparameters can yield different results; therefore, extension to longitudinal datasets is not straightforward, unlike traditional algorithms such as PCA. Aligned-UMAP is a recently introduced dimensionality reduction approach for temporal data by the authors of UMAP (`https://umap-learn.readthedocs.io/en/latest/aligned_umap_basic_usage.html`). It is based on the UMAP [84] and MAPPER [87] algorithms. MAPPER is a well-known topological data analysis method that successfully studies temporal, unbiased transcriptional regulation patterns [88]. Aligned-UMAP imposes time constraints in the low dimensional embeddings, thereby controlling the stochasticity of its cross-sectional counterpart along the longitudinal axis. TimeCluster [89] is another approach that reduces the dimensionality of time-series data. Though it is possible to discover clusters with similar trajectories using TimeCluster, their intrinsic longitudinal variation cannot be observed. Further, it requires data availability for every time instance, making it less applicable for most biological datasets.

In this work, we deep dive into the applications of Aligned-UMAP on various longitudinal biological datasets. We applied the algorithm to clinical data, brain images, longitudinal proteomic data, EHR, and ECG datasets. We demonstrated its utility for researchers to identify exciting patterns and trajectories within enormous data sets. Secondly, we show the effect of different parameters of Aligned-UMAP on the lower dimension space. We also performed computation time analysis with varying datasets as a factor of the number of CPU cores. Furthermore, we deployed an interactive data visualization tool for reproducibility and transparency, motivated by open science. A deeper investigation of observed patterns



Method1: UMAP

- Apply UMAP on all samples ignoring time dependency structure
- Visualization using 2D scatter plots

Method2: Aligned UMAP

- Apply UMAP at each time step
- Regularize embedding by imposing time constraints
- Visualization using 3D line plots

Figure 3.2: The workflow of analysis and model development.

could reveal more detailed, meaningful information, which is out of the scope of this work.

3.2 METHODS

3.2.1 Data pre-processing

All the datasets went through data processing before applying the Aligned-UMAP algorithm. We follow the same methodology used in the cited publications (Table 1). Here we list the summary of the data processing details for each of the datasets used in this work:

- PPMI clinical data [32, 49]: This clinical data was obtained from the Parkinson's Progression Marker Initiative (PPMI, `http://www.ppmi-info.org/`). Data went through triage for missing data, a 60-month assessment, and comprehensive phenotype collection. In the study, we included only data from participants with 60 months of follow-up for PPMI. Overall, in the PPMI (n = 294 PD cases including 99 (34%) female; 154 controls including 58 (38%) female) passed the triage. We color the trajectory based on progression-based subtypes obtained from Dadu et al. [49]. We used the source code located at `https://github.com/anant-dadu/PDProgressionSubtypes`.

- ADNI clinical data [90]: Clinical assessment data for Alzheimer's disease were obtained from the ADNI database (`https://adni.loni.usc.edu/`). The total scores and subscores from commonly collected cognitive, functional, and longitudinal clinical data elements were aggregated to form a 78-dimension feature vector. Missing values were estimated using linear interpolation based on the past visit readings for the feature, avoiding any influence of other observations during data imputation as per Satone et al. [58]. For our analysis, we utilized the code provided at `https://github.com/NIH-CARD/ADProgressionSubtypes`.

- PPMI-ADNI T1 MRI [91]: In this dataset, we used derived features that include regional brain volumes, cortical thickness, and area as T1 MRI imaging features. We used ANTsPyT1w available at `https://github.com/stnava/ANTsPyT1wtopre-processtheimages`.

- MIMIC-III [92]: We utilized the data processing code available at `https://github.com/Jeffreylin0925/MIMIC-III_ICU_Readmission_Analysis` to generate features from electronic health records. We used three categories of features in this work, namely chart events, ICD-9 embeddings, and demographic information of the patients [93].

We download the preprocessed version for the other three datasets using the link provided in the relevant publications, Longitudinal Proteomic COVID–19 from Filbin et al. [94], Longitudinal whole lung scRNA from Strunz et al. [95] and iPSC derived neurons from Reilly et al. [96]. On all these datasets, we applied min–max normalization to numerical features to preserve the longitudinal relationships among the original data and ensure a zero-to-one range. Additionally, we outlined the specifics of data preparation in the Readme file of our publicly accessible GitHub repository (`https://github.com/NIH--CARD/AlignedUMAP--BiomedicalData#step1-prepare--data`).

3.2.2 Statistical, and machine learning analysis

After preparing the data, we perform unsupervised machine learning using the Aligned-UMAP algorithm. We hypothesized that this approach could identify the clusters with distinct trajectories over time. Since this work is an entirely unsupervised analysis, we visualize 3D trajectory plots, color-coded based on metadata, to evaluate the algorithm's performance. We performed extensive hyperparameter tuning with different sets of values for Aligned-UMAP parameters (distance metric, alignment regularization, alignment window size, number of neighbors, minimum distance). For additional information, please see Section 2 of the Readme file available in our GitHub repository at (`https://github.com/NIH-CARD/AlignedUMAP-BiomedicalData#step2-setup-configuration-and-data-paths`). Finally, we analyze the time taken by Aligned-UMAP on all our datasets to provide the estimate of execution time to the users (Figure 3.11).

3.3 RESULTS

3.3.1 Overview of the Aligned-UMAP method

Uniform Manifold Approximation and Projection: UMAP is a dimensionality reduction method that learns a non-linear low-dimensional embedding of the original high-dimensional space. UMAP has solid theoretical foundations based on manifold theory and tries to preserve both the local and some global structure better than other popular techniques such as t-SNE. UMAP is a graph-based dimensionality reduction method. It has two phases—first, computation of a weighted nearest-neighbor graph from the high-dimensional dataset. In the second phase, a low-dimensional layout is computed by optimizing the objective function that preserves desired characteristics of this nearest-neighbor graph. The algorithm is computationally efficient with the time order of sample size for the low dimensional optimization phase but is essentially bounded by the log-linear complexity of the nearest neighbor search phase in practical scenarios [97]. It is the superior run time performance of UMAP as compared to its counterparts that makes it very popular among the dimensionality reduction methods [85].

Aligned Uniform Manifold Approximation and Projection: Aligned-UMAP is a recently introduced dimensionality reduction approach for temporal data. The trivial way of performing dimensionality reduction on longitudinal data is to apply UMAP independently at different time steps and align the embedding using a Procrustes transformation on related points. However, Aligned-UMAP optimizes both embeddings simultaneously using a regularizer term to provide better alignments in general. The MAPPER algorithm is used to get the regularizer term which enforces the constraint on how far related points can take different locations in embeddings at multiple time points. Further details for the algorithm can be found on the UMAP documentation website (`https://umap-learn.readthedocs.io/en/latest/aligned_umap_basic_usage.html`). Figure 3.2 and Figure 3.1 shows the pipeline of our analysis workflow.

3.3.2 Software output and reproducibility

A demo of the Aligned-UMAP visualization is available at `https://alignedumap-biomedicaldata.streamlit.app`. The data analysis pipeline for this work was performed in Python 3.8 using open–source libraries (numpy, pandas, plotly, umap). Our code is publicly available at `https://github.com/NIH-CARD/AlignedUMAP-BiomedicalData` to facilitate replication and future expansion of our work. The repository is well documented and includes a description of the data pre-processing, statistical, and machine learning analysis used in this study.

3.3.3 Visualizing high-dimensional longitudinal data

We study Aligned-UMAP in a wide range of biomedical datasets from multiple data modalities. Table 3.1 shows the statistics of various datasets, with the count of samples ranging from approximately 500 to 21,000. These datasets vary in both the number of time sequences and the number of features available. For every visualization, each representative point becomes a thread through the time-axis as their relative position changes in the low-dimensional space. Low dimensional embeddings by UMAP and Aligned-UMAP dimensionality reduction algorithms on longitudinal biomedical datasets from multiple modalities are shown from Figure 3.3- 3.9. Also, it should be noted that we apply the Aligned-UMAP algorithm on the dataset having characteristics shown in Table 3.1. In this figure, we have demonstrated a subset of classes for better visualization purposes. For more detailed analysis, users can explore our public web application.

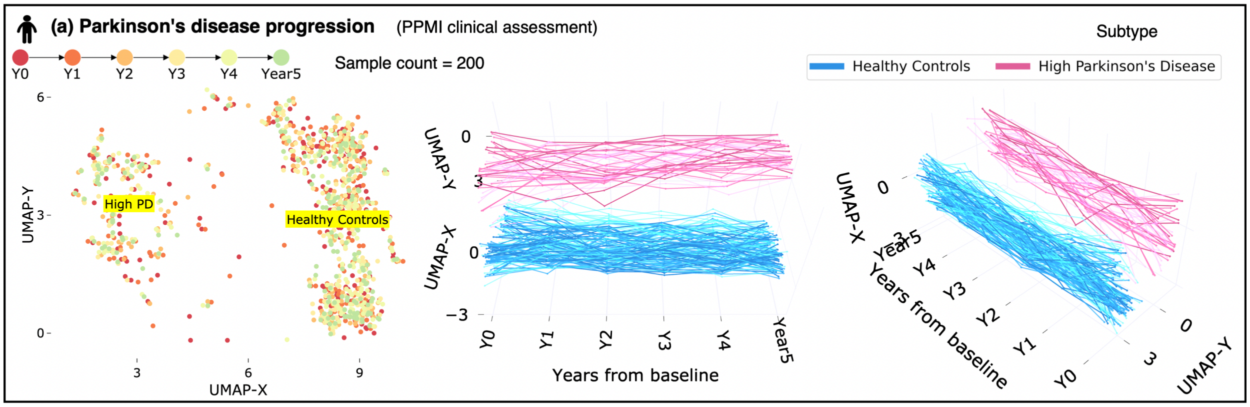


Figure 3.3: Reveals the distinction between Parkinson's Disease subjects (with rapid progressors) and Healthy Controls from 122 clinical measurements collected over five years from Parkinson's Progression Markers Initiative (PPMI) study. Measures include Montreal Cognitive Assessment scores and MDS-Unified Parkinson's Disease Rating Scale scores.

Clinical data: In neurodegenerative diseases such as Alzheimer's and Parkinson's, the individual can manifest disease in various ways, often times prior to clinical diagnosis. We evaluate the Aligned-UMAP algorithm on the clinical assessment data from Alzheimer's Disease Neuroimaging Initiative (ADNI) and Parkinson's Progression Markers Initiative (PPMI) study cohorts. The ADNI study includes Alzheimer's patients, mild cognitive impairment subjects, and elderly controls. PPMI study has subjects recently diagnosed with Parkinson's

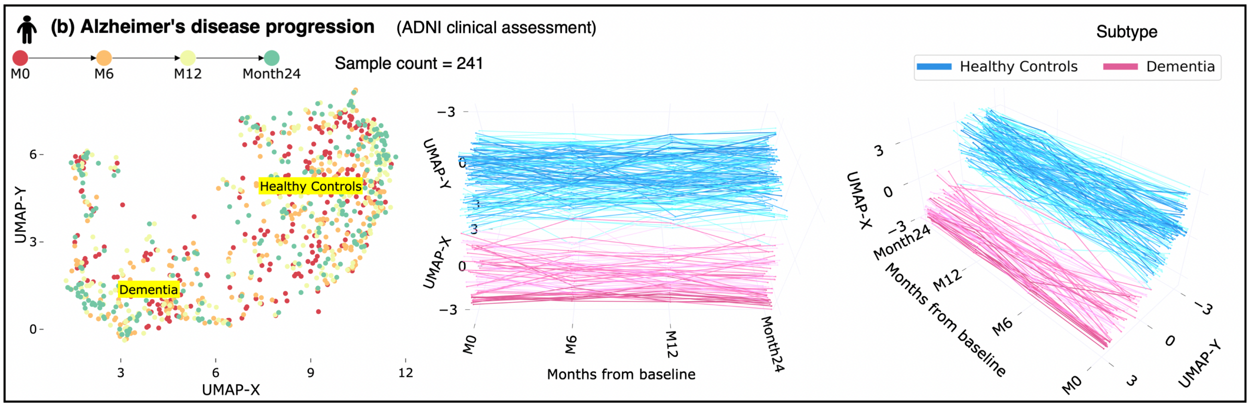


Figure 3.4: Show trajectories of Dementia and Healthy Control subjects on 78 clinical measurements collected over two years from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. Measurements include Mini-Mental State Exam (MMSE) scores and Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-COG) tests.

Disease (PD) and healthy controls. These studies collect data for many clinical assessments related to movement and cognitive disability to monitor disease progression. All such measurements are recorded longitudinally at separate visits. The time duration of such visits can range from years to decades.

We preprocess the ADNI and PPMI cohort datasets following the strategy proposed in previous disease subtyping studies [49, 58, 98]. UMAP and Aligned-UMAP successfully pulled together clusters corresponding to populations with similar disease progression (Figure 3.3 and Figure 3.4). However, longitudinal differences got lost in the UMAP version due to its stochastic nature. Aligned-UMAP separates rapidly progressive PD from the healthy control group and demonstrates divergence of the rapid PD subgroup from healthy controls with aging (Figure 3.3). Furthermore, Aligned-UMAP reveals distinct longitudinal courses for dementia and the healthy control group (Figure 3.4). We follow a continuum spectrum from lower progressive to high progressive subgroups for PD and Dementia subjects. These results suggest that Aligned-UMAP could be used as a hypothesis-generating tool to identify distinct subtypes based on disease progression. For instance, a particular subgroup shows rapid decline in clinical symptoms such as MDS-Unified Parkinson's Disease Rating Scale [61] or MoCA cognitive assessment [62] as compared to healthy control and other subgroups.

Whole lung scRNA Data: Single-cell transcriptomics (scRNA) using next-generation transcript sequencing (RNA-seq) has recently received much attention due to its ability to uncover cellular heterogeneity, cellular differentiation, and development mechanisms. UMAP has demonstrated its efficacy in analyzing single-cell datasets by identifying clusters of related cells. Modeling gene expression trajectories of different cell types have been successfully used

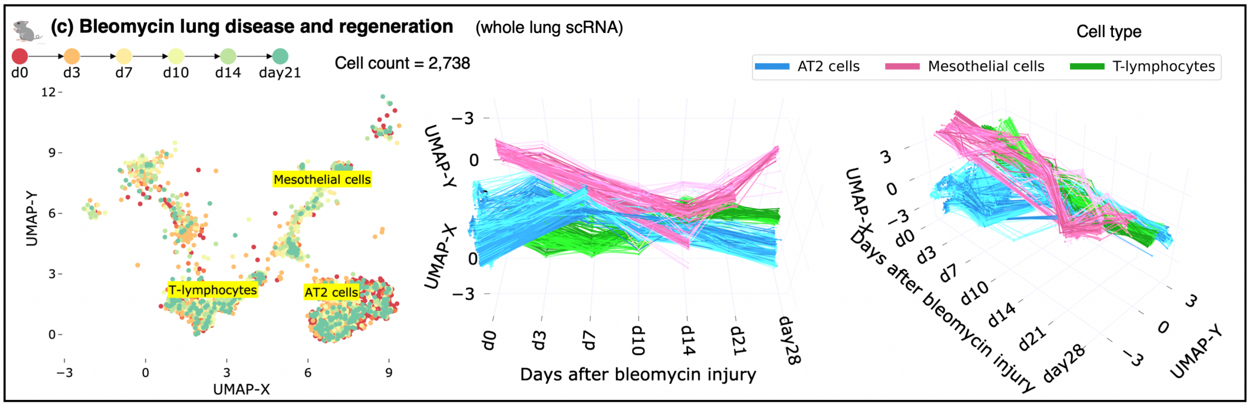


Figure 3.5: Aligned-UMAP trajectories show shifts in specific cell types (such as Mesothelial and AT2 cells) in gene expression space during the regeneration time course of mice having bleomycin lung injury.

to understand cell-cell communication routes in various chronic diseases such as lung disease and tumor cells [95, 99]. We evaluated Aligned-UMAP on whole lung scRNA data of mice undergoing regeneration after bleomycin-induced lung injury [95]. Transcriptomic profile of 29,297 cells was collected from six time points (day 3, 7, 10, 14, 21, and 28). We observe clusters of cell types showing different cellular dynamics through the regeneration process (Figure 3.5); Mesothelial cells show a spike at day 14 and start returning to their healthy state (day 0). Thereby, suggesting the role of mesothelial cells in bleomycin related lung injury. This way, we could extract hidden longitudinal patterns from high-dimensional time-series datasets using Aligned-UMAP.

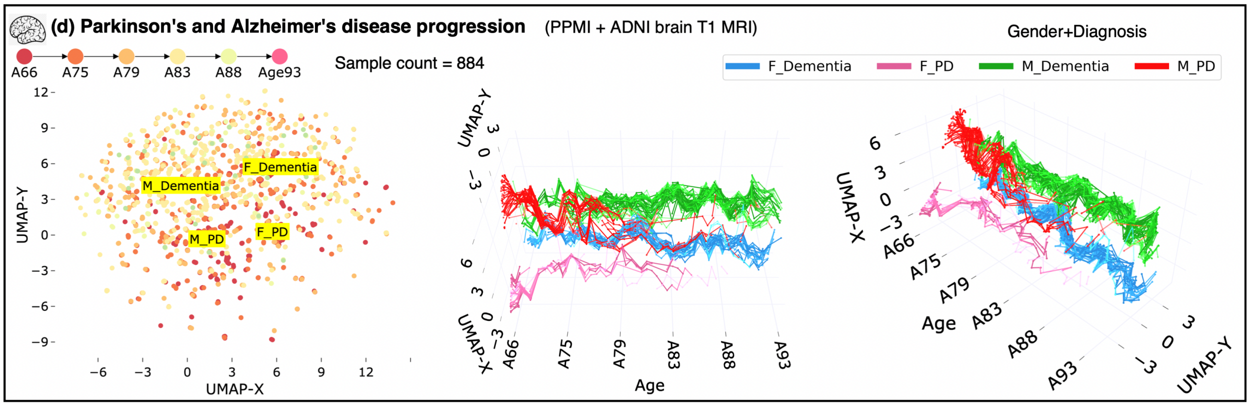


Figure 3.6: Aligned-UMAP embeddings depict aging patterns for Dementia, and Parkinson's disease patients, stratified by gender.

Imaging Data: Imaging is a pervasive way of monitoring the disease progression of multiple disorders. We use the Advanced Normalization Tools (ANTs) pipeline 18 to extract structural features such as the volume and area of different brain regions from the Magnetic Resonance Imaging (MRI) T1 image. Since the number of longitudinal images for each subject is scarce, we use the imaging features to model aging trajectories. To be precise, we relate images if they are observed at similar age groups instead of relating subjects based on their visits. Also, these relations are constrained by different diagnosis groups (i.e., Control, PD, or Dementia). Figure 3.6 shows various aging courses based on the subject's latest diagnosis and gender. We noticed a more rapid decline among female dementia cases versus male dementia cases around 80 years of age. It suggests the non-linear and distinct patterns of disease progression across groups within a disease. We observed distinct longitudinal trajectory patterns, which might be a possible way to monitor disease progression.

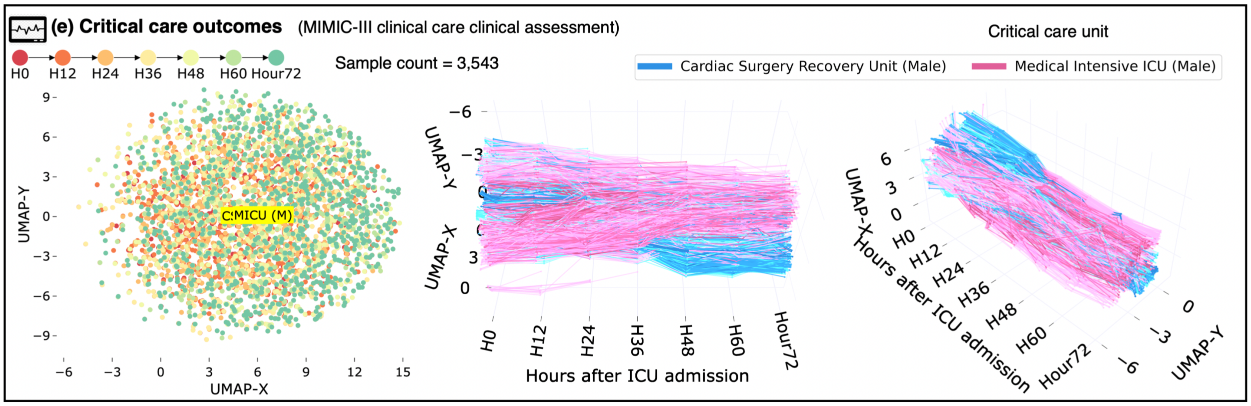


Figure 3.7: Unveils trajectories of the subject's admitted in different critical care units of the MIMIC-III database. Measurements include vital signs such as blood pressure, oxygen levels, and ICD-9 diagnosis codes.

EHR Data: Electronic Health Records (EHR) is a systematic collection of patients' healthcare records in a digital format. EHR is adopted in many hospitals in the USA and UK [100]. We applied the Aligned-UMAP on the MIMIC-III Critical Care Database [92], which consists of records of more than 40,000 patients in intensive care units (ICU) of the Beth Israel Deaconess Medical Center between 2001 and 2012. We preprocessed the dataset following the methodology proposed by Lin et al. [93]. Figure 3.7 shows the lower-dimensional space on the MIMIC-III dataset on measurement recorded during the initial 72 hours. of entry to the Intensive Care Unit (ICU). We color the trajectories based on the type of critical care unit a patient stays in just before discharge from the hospital. We observe that UMAP could not recover time-related patterns; however, Aligned-UMAP segregates trajectories based on the patient's critical care unit. This pattern reflects that it might be helpful to analyze ICU datasets stratified by their care unit and suggests that the quality of care in ICU units is highly variable.

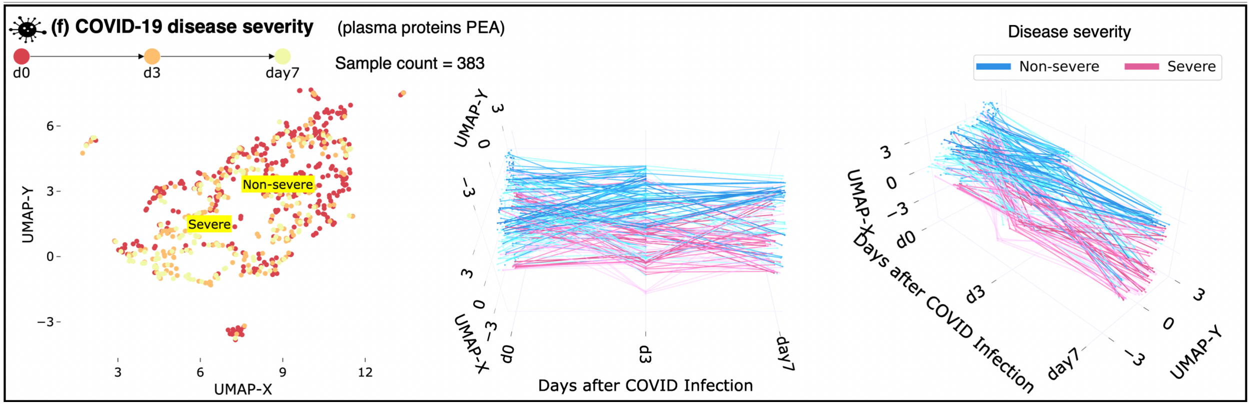


Figure 3.8: Embedding space depicts the severity of COVID-19 disease from 1,463 unique plasma proteins measured by proximity extension assay using the Olink platform. The cutoff at day 3 is visible because of data unavailability at day seven due to either patient recovery or deceased

COVID-19 Proteomics Data: Uncovering protein signatures associated with COVID-19 infection and severity can provide insights into its pathophysiology and immune dysfunction [94]. We utilized longitudinal proteomic data on 306 COVID-19 patients [94]. Aligned-UMAP has identified distinct trajectories for severe and non-severe patients over seven days (Figure 3.8). We observed the participants exhibiting continued negative symptom trajectories at seven days belonging to more severe or longer COVID-19 infection.

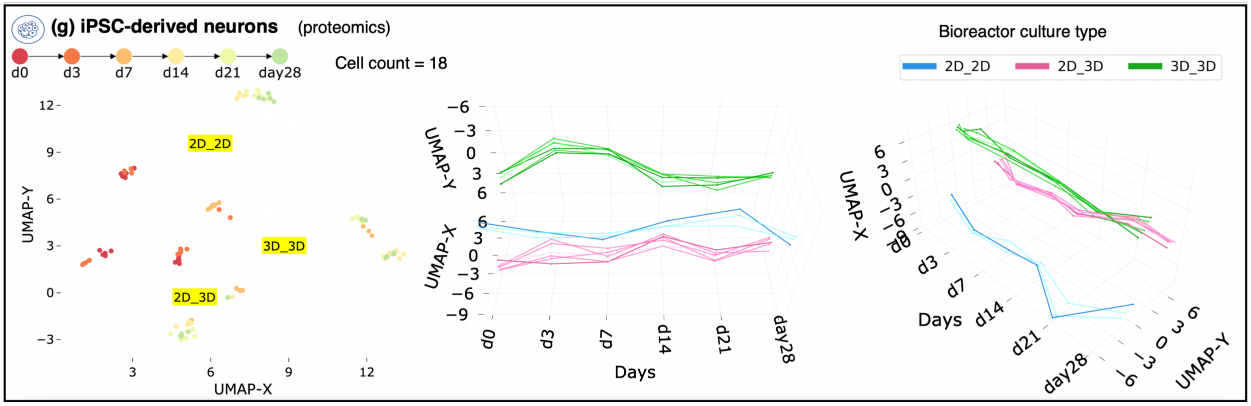


Figure 3.9: Aligned-UMAP low dimensional space identified the cell culture environment of iPSC-derived neurons using longitudinal proteomic data for more than 8000 proteins.

iPSC-derived neurons Proteomics Data: Aligned-UMAP can be incorporated as a quality control measure for longitudinal data. We applied this approach to longitudinal proteomic profiling of the differentiation of iPSC (Induced Pluripotent Stem Cells) derived neurons cultured in different bioreactors [96]. We could visualize distinct patterns of change for each cell line grouped by their culture environment, thereby identifying batch effects

(Figure 3.9). We observed that the cell lines cultured only in the 2D bioreactor are hypervariable for almost all time points (till day 28). The cell line 2D_3D (day0-day3 2D culture, day4-day28 3D culture) tends to converge around day 14, and the cell line cultured in the 3D bioreactor tends to be more homogeneous after around day 7. A tighter spread denotes a homogeneous group.

Table 3.1: Datasets overview and statistics

Dataset	Modality	# samples	# features	# time sequences
PPMI clinical data	Clinical Assessment	476	122	6
ADNI clinical data	Clinical Assessment	435	78	4
PPMI-ADNI T1 MRI	MRI T1 Imaging	2,836	406	52
MIMIC-III	EHR	36,675	64	6
Longitudinal Proteomic COVID-19	Proteomics	383	1,463	3
Longitudinal whole lung scRNA	scRNA	10,111	21,767	7
iPSC derived neurons	Proteomics	18	4,959	6

3.4 DISCUSSION

3.4.1 Observed meaningful patterns

Our work demonstrates that Aligned-UMAP could help us discover meaningful longitudinal patterns by color-coding them based on multiple known covariates. Our analysis finds that both UMAP and Aligned-UMAP help generate intuitive embeddings because of their ability to preserve the global structure. Additionally, Aligned-UMAP provides a view that highlights longitudinal structure by imposing time constraints in the embeddings, thereby controlling the stochasticity of its cross-sectional counterpart. We observe distinct trajectory patterns of the data from different modalities. Dementia and PD subtypes are delineated using clinical assessment measurements from the PPMI and ADNI study (Figure 3.3, Figure 3.4). Aligned-UMAP has also shown visually meaningful patterns on high-dimensional omics data such as proteomics (Figure 3.8, Figure 3.9) or single-cell transcriptomics data (Figure 3.5). Therefore, it is evident that Aligned-UMAP provides meaningful representations and is likely to be a valuable tool for researchers working on multivariate longitudinal datasets by preserving the global and local trends along the time axis.

3.4.2 Points to remember

Based on our observations from this study, this approach promises to be useful in many other biomedical datasets. These datasets can vary in terms of missing data, time sequences, or domain-specific variations that make it challenging to tune experimental settings. So, here we discuss key points that users should keep in mind while using Aligned-UMAP.

- **Missing data:** The problem of missing data is prevalent in healthcare datasets and can interfere with the conclusions drawn from the data. Aligned-UMAP can handle data missingness across the longitudinal dimension by performing interpolation in low-dimensional space. Tensor decomposition based dimension reduction approaches cannot handle any data missingness [89]. However, none of the dimension reduction approaches are designed to handle missingness for features measured cross-sectionally.

- **Aligned-UMAP Parameter Effect:** The number of neighbors and the minimum distance are two critical parameters affecting the lower-dimensional space using the UMAP algorithm. In Aligned-UMAP, the number of parameters can increase significantly. We can vary the UMAP parameters for each step to observe different trajectories. The two other alignment parameters, namely, alignment window size and alignment regularizer, are critical in visualizing the longitudinal trend that controls the volatility along the time axis. Figure 3.10 shows the effect of alignment window size, alignment regularizer, and the number of neighbors on the PPMI longitudinal dataset. Our web app also demonstrates the impact of these parameters on the lower-dimensional space.

- **Execution Time:** We analyze the execution time taken by both algorithms on multiple datasets and use their subsamples of different sizes. Further, to understand the algorithm's scalability and parallelization, we executed it utilizing different numbers of cores (Figure 3.11). Multiple core setup does not seem to improve run times of Aligned-UMAP in low data regimes, which may be attributed to inter-core synchronization overheads. However, significant improvements are observed on complete lung scRNA data with 16 cores (Figure 3.11**a**). Compared to UMAP, Aligned-UMAP would require a larger dataset to have better parallelization on a multi-core machine (Figure 3.11**b**).

- **Stochastic models and reproducibility:** Although Aligned-UMAP can handle stochasticity along the longitudinal axis, it still produces variable embeddings on different runs. Like UMAP, it uses randomness both to speed up approximation steps and to aid in solving optimization problems, thereby affecting the reproducibility of the lower dimensional space. However, UMAP and Aligned-UMAP provide relatively stable results when applied to large amounts of data. In the future, sophisticated approaches are required to ensure reproducibility.

3.5 FUTURE WORK

The Aligned-UMAP algorithm is still in the developing phase. We discuss the plausible extensions of the algorithm that might be useful in a multitude of biomedical research datasets.

- Clustering: The dimensionality reduction method is a standard pre-processing step to utilize density-based clustering methods on the high-dimensional dataset. Dynamic time warping is the most common metric to cluster time-varying patterns using K-mean clustering. It will be interesting to evaluate multiple clustering approaches on longitudinal trajectories.

- Semi-Supervised / Supervised: Sometimes, we would like to incorporate target label information to project high-dimensional data to lower-dimensional space in dimensionality reduction. There are various reasons for supervised dimension reduction; First, to retain the internal structure of classes and have dense clusters. Secondly, to maintain the global structure. i.e., preservation of inter-relationships among the known classes. Finally, we can observe outliers or subjects that do not belong to either class using the semi-supervised learning approach. The extension of Aligned-UMAP for supervised/semi-supervised dimension reduction will be a part of future work.

- Rare Events Detection: UMAP algorithm supports the detection of outliers using the Local Outlier Factor [101] algorithm. Identifying outliers from longitudinal trajectories generated by Aligned-UMAP will need further investigation.

- Multi-modal aspect: In the biomedical domain, monitoring disease needs data from multiple modalities such as imaging, blood biomarkers, genetics, or multi-omics [53, 54]. Current dimensionality reduction approaches are destined for the dataset from a single modality. The trivial way of incorporating multi-modal data is to use vectorization, but it might not be the optimal solution to discover hidden patterns in the data. Therefore, evaluating and building new dimensionality reduction approaches for multi-modal data analysis setup is required.

- Interpretability: It's important to note that because UMAP and t-SNE both necessarily warp the high-dimensional shape of the data when projecting to lower dimensions, any given axis or distance in lower dimensions still isn't directly interpretable in the way of techniques such as PCA. However, PCA is highly influenced by outliers present in the data, and its inability to capture nonlinear dependencies causes a mix up among underlying clusters in lower dimensional space.

- Data frequency: Since Aligned-UMAP creates a lower-dimensional space for every location, analyzing data collected at an extremely fine scale, such as ICU or ECG spectrograms, becomes expensive.

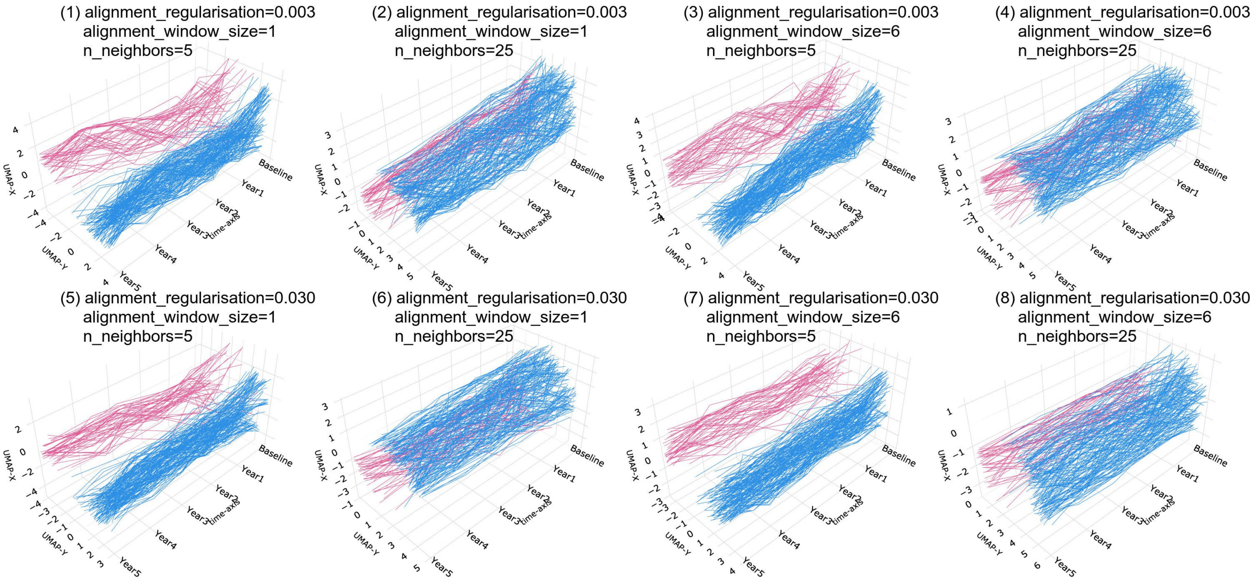


Figure 3.10: Effect of hyperparameters of Aligned-UMAP on the PPMI clinical dataset. The alignment regularization is varied for [0.003, 0.03], alignment window size from [1, 6] and number of neighbors from [5, 25]. We could observe that an increase in the number of neighbors increases the size of visible clusters (1, 2). Alignment regularization controls and alignment window size the volatility of trajectories. Higher values for alignment regularization will keep the related embeddings closer (1, 5), and alignment window size captures how far forward and backward across the datasets we look at when doing alignment (1, 3).

So far, we have seen the potential of unsupervised learning in detecting patterns in an unbiased way. In the next chapter, we will explore how we can utilize labeled data. Supervised algorithms are also useful when we have labels to guide the algorithms towards specific tasks. Furthermore, we have focused on utilizing brain imaging and genomic data to compare different NDDs, which allows for early detection during the asymptomatic phase.

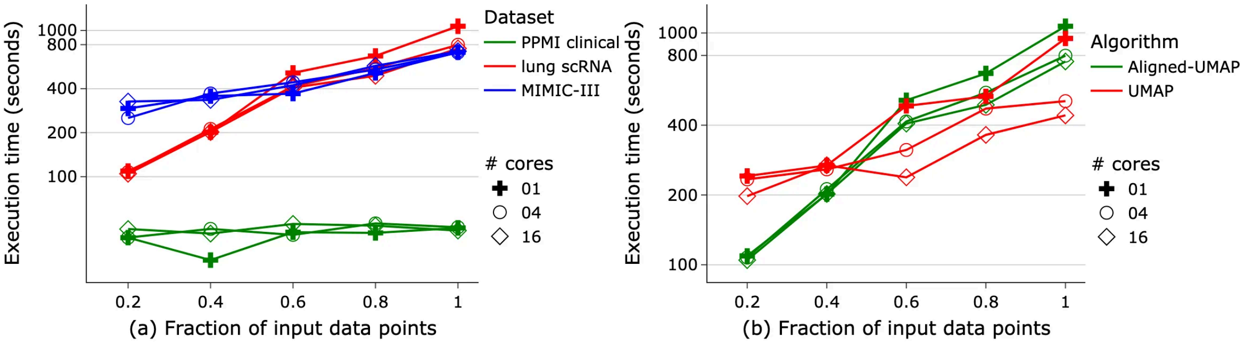


Figure 3.11: Execution time for input datasets of varying sizes (a) for Aligned-UMAP on multiple datasets (b) for Aligned-UMAP and UMAP on whole lung scRNA dataset. All the experiments are conducted on a 128 GB RAM machine utilizing a different number of cores (marker symbol).

CHAPTER 4: PREDICTION, PROGNOSIS AND MONITORING OF NEURODEGENERATION AT BIOBANK-SCALE

Alzheimer's disease and related dementias (ADRD) and Parkinson's disease (PD) are the most common neurodegenerative disorders. These are multisystem disorders affecting different body parts and functions. Patients with ADRD or PD have long asymptomatic phases and exhibit significant etiology or clinical manifestations heterogeneity. Hence, quantitative measures that can provide early disease indicators are necessary to improve patient stratification, clinical care, and clinical trial design. This work uses machine learning techniques to derive such a quantitative marker from T1-weighted (T1w) brain Magnetic Resonance Imaging (MRI). In this retrospective study, we developed a machine learning (ML) based score of T1w brain MRI image utilizing disease-specific Parkinson's Disease Progression Marker Initiative (PPMI) and Alzheimer's Disease Neuroimaging Initiative (ADNI) cohorts. Then, we evaluated the potential of ML-based scores for early diagnosis, prognosis, and monitoring of ADRD and PD in an independent large-scale population-based cohort, UK Biobank, using longitudinal data. In this analysis, 1,826 dementia (from 731 participants), 3,161 healthy controls images (925 participants) from the ADNI cohort, 684 PD (319 participants), 232 healthy controls (145 participants) from PPMI cohort were used to train machine learning models. The classification performance is 0.94 [95% CI: 0.93-0.96] area under the ROC Curve (AUC) for ADRD detection and 0.63 [95% CI: 0.57-0.71] for PD detection using 790 extracted structural brain features. We identified the hippocampus and temporal brain regions as significantly affected by ADRD and the substantia nigra region by PD. The normalized ML model's probabilistic output (ADRD and PD imaging scores) was evaluated on 42,835 participants with imaging data from UK Biobank. For diagnosis occurrence events within 5 years, the integrated survival model achieves a time-dependent AUC of 0.86 [95% CI: 0.80-0.92] for dementia and 0.89 [95% CI: 0.85-0.94] for PD. ADRD imaging score is strongly associated with dementia free survival (hazard ratio (HR) 1.76 [95% CI: 1.50-2.05] per S.D. of imaging score), and PD imaging score shows association with PD free survival (hazard ratio 2.33 [95% CI: 1.55-3.50]) in our integrated model. HR and prevalence increased stepwise over imaging score quartiles for PD, demonstrating heterogeneity. The scores are associated with multiple clinical assessments such as Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog), and pathological markers, including Amyloid and Tau. Finally, imaging scores are associated with polygenic risk scores for multiple diseases. Our results indicate that we can use imaging scores to assess the genetic architecture of such disorders in the future. Our study demonstrates the use of quantitative markers generated using machine learning techniques for ADRD and PD. We show that disease probability scores obtained from brain structural features are useful for early detection, prognosis prediction, and monitoring disease progression.

4.1 INTRODUCTION

Neurodegeneration refers to the progressive loss of neurons, causing the loss of brain functions. Alzheimer's disease and related dementias (ADRD) and Parkinson's disease (PD) are the most common forms of neurodegeneration affecting millions of people worldwide [13, 14, 102]. ADRD is predominantly characterized as a cognitive or behavioral disorder, while PD mainly affects motor skills [11]. However, these are multisystem disorders affecting different body parts and functions. Further, both these diseases exhibit substantial phenotypic heterogeneity with clinical manifestations varying by onset age, progression rate, or the constellation of motor/non-motor features [23, 49, 58, 103, 104]. ADRD and PD are now recognised as forming a phenotypic spectrum rather than discrete categories. Therefore, the binary assignment cannot capture the complexity of ADRD and PD. Hence, quantitative indicators of disease are required. Such quantitative markers can also function as surrogate endpoints to increase clinical trials' ability to monitor treatment effects [105].

Clinical diagnosis of neurodegenerative disease is preceded by a potentially long asymptomatic phase [106]. It is critical to identify the disease during patients' pre-asymptomatic phase when there are higher chances of successful disease modifying therapies [107]. Identifying the disease during patients' pre-asymptomatic phase is critical when there are higher chances of successful disease modifying therapies. There has been evidence of abnormal anatomical changes during the pre-asymptomatic phase of AD and PD due to neuronal loss [108]. Further, the risk of getting these disorders is highly dependent on heritable factors, with more than 75 AD and 90 PD associated genetic risk loci already identified [38, 39]. Together, both genetic and imaging data modalities could be a useful predictor of clinical outcomes during the asymptomatic period of neurodegeneration.

Machine learning approaches have facilitated the analysis of a large number of brain structural features from T1-weighted (T1w) Magnetic resonance imaging (MRI) [109]. Machine learning models have shown success in accurately diagnosing AD (post-clinical diagnosis) [110, 111], as well as predicting the conversion from mild cognitive impairment to AD (pre clinical diagnosis) [112, 113]. These approaches have been primarily used as a classification tool to discriminate patients from healthy individuals (binary classification), rather than quantifying disease. Further, very few studies have validated their findings in an independent cohort [114]. While T1-weighted MRI is commonly used to study AD, its application in PD is limited due to the need for more fine-grained features to analyze neuronal loss in the substantia nigra, which is the area most affected in PD [115].

In the following, we present how we developed quantitative markers of ADRD and PD from brain imaging data using machine learning techniques and investigated their association with clinical outcomes during the pre-diagnosis and post-diagnosis phases. We evaluated the contribution of imaging scores, in combination with genetic risk factors to predict the likelihood of developing the disease later in life. In the post-diagnosis phase, we assessed the potential of imaging scores as a disease monitoring marker by examining their association with clinical assessments and relevant pathological biomarkers. We leverage disease-specific cohorts to train the machine learning models and evaluate machine learning-generated imaging scores in a large external biobank. The cost-effective biomarker using T1 brain imaging could pave the way for passive surveillance of the healthcare system for neurodegenerative disorders at a biobank scale. Figure 4.1 highlights the workflow for our analysis.

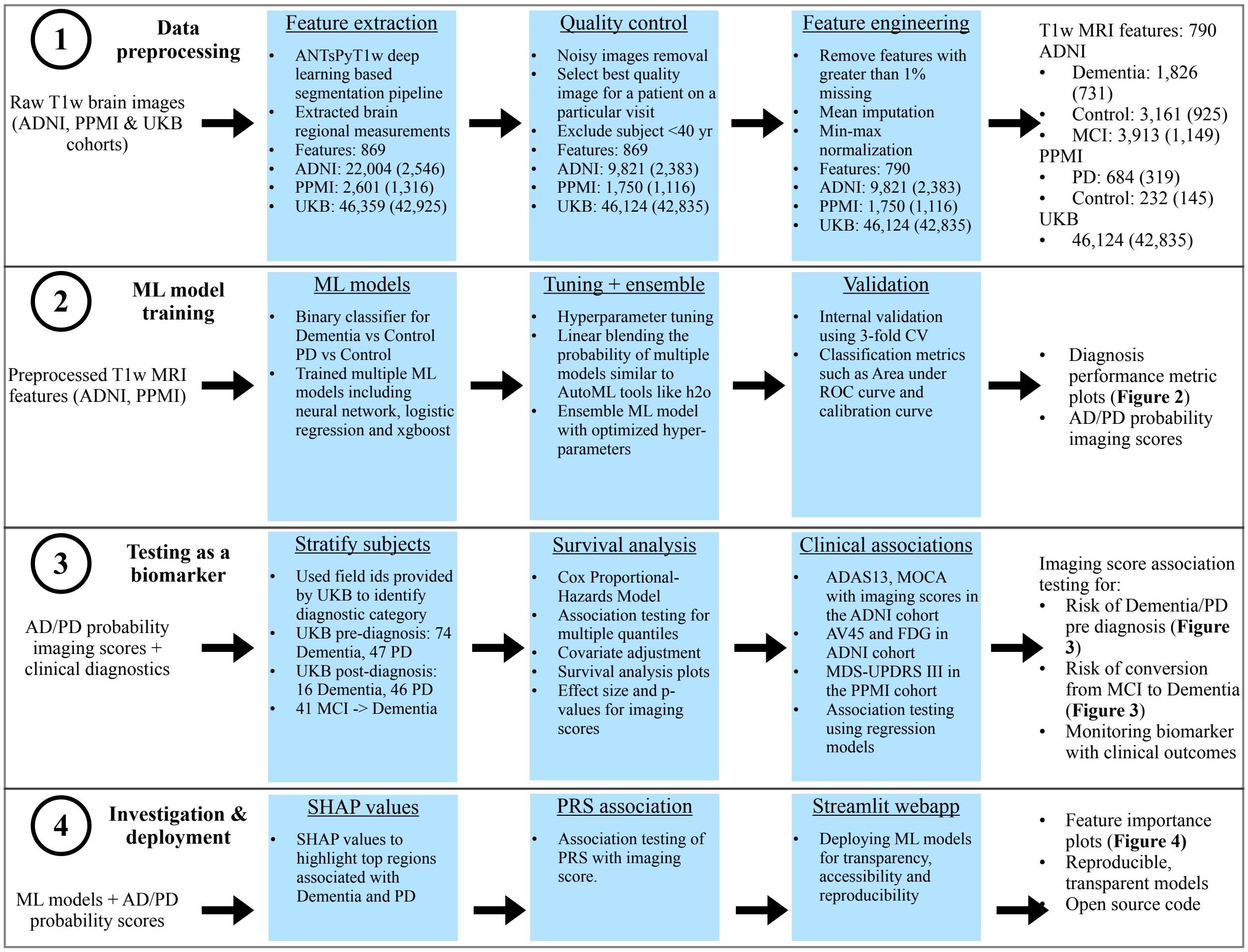

1
Data preprocessing
Raw T1w brain images (ADNI, PPMI & UKB cohorts)
Feature extraction
• ANTsPyT1w deep learning based segmentation pipeline
• Extracted brain regional measurements
• Features: 869
• ADNI: 22,004 (2,546)
• PPMI: 2,601 (1,316)
• UKB: 46,359 (42,925)
Quality control
• Noisy images removal
• Select best quality image for a patient on a particular visit
• Exclude subject <40 yr
• Features: 869
• ADNI: 9,821 (2,383)
• PPMI: 1,750 (1,116)
• UKB: 46,124 (42,835)
Feature engineering
• Remove features with greater than 1% missing
• Mean imputation
• Min-max normalization
• Features: 790
• ADNI: 9,821 (2,383)
• PPMI: 1,750 (1,116)
• UKB: 46,124 (42,835)
T1w MRI features: 790
ADNI
• Dementia: 1,826 (731)
• Control: 3,161 (925)
• MCI: 3,913 (1,149)
PPMI
• PD: 684 (319)
• Control: 232 (145)
UKB
• 46,124 (42,835)
2
ML model training
Preprocessed T1w MRI features (ADNI, PPMI)
ML models
• Binary classifier for Dementia vs Control PD vs Control
• Trained multiple ML models including neural network, logistic regression and xgboost
Tuning + ensemble
• Hyperparameter tuning
• Linear blending the probability of multiple models similar to AutoML tools like h2o
• Ensemble ML model with optimized hyper-parameters
Validation
• Internal validation using 3-fold CV
• Classification metrics such as Area under ROC curve and calibration curve
• Diagnosis performance metric plots (Figure 2)
• AD/PD probability imaging scores
3
Testing as a biomarker
AD/PD probability imaging scores + clinical diagnostics
Stratify subjects
• Used field ids provided by UKB to identify diagnostic category
• UKB pre-diagnosis: 74 Dementia, 47 PD
• UKB post-diagnosis: 16 Dementia, 46 PD
• 41 MCI -> Dementia
Survival analysis
• Cox Proportional-Hazards Model
• Association testing for multiple quantiles
• Covariate adjustment
• Survival analysis plots
• Effect size and p-values for imaging scores
Clinical associations
• ADAS13, MOCA with imaging scores in the ADNI cohort
• AV45 and FDG in ADNI cohort
• MDS-UPDRS III in the PPMI cohort
• Association testing using regression models
Imaging score association testing for:
• Risk of Dementia/PD pre diagnosis (Figure 3)
• Risk of conversion from MCI to Dementia (Figure 3)
• Monitoring biomarker with clinical outcomes
4
Investigation & deployment
ML models + AD/PD probability scores
SHAP values
• SHAP values to highlight top regions associated with Dementia and PD
PRS association
• Association testing of PRS with imaging score.
Streamlit webapp
• Deploying ML models for transparency, accessibility and reproducibility
• Feature importance plots (Figure 4)
• Reproducible, transparent models
Open source code

Figure 4.1: Workflow of analysis and model development.

4.2 RESEARCH IN CONTEXT

4.2.1 Evidence before this study

We searched PubMed for articles published in English from database inception to May 11, 2023, about the use of machine learning on brain imaging data for Alzheimer's disease (AD), dementia, and Parkinson's disease (PD) population. We used search terms "machine learning" AND "brain imaging" AND "neurodegenerative disorders" AND "quantitative biomarkers". The search identified 25 studies. Most of these studies are focused on Alzheimer's disease. They use machine learning to predict conversion from mild-cognitive impairment to dementia or to build a classification tool. Many studies also focused on positron emission tomography (PET) images rather than cost effective T1w MRI images in their analysis. None of the studies have focused on detecting disease during the asymptomatic phase of dementia and PD. Identified studies are limited in sample size (order of hundred samples) and extracted features. The assessments of the clinical utility of machine learning predicted disease probabilities are scarce. Significantly, no attempts were made to validate the algorithm in an external cohort.

4.2.2 Added value of this study

This study developed machine learning based quantitative scores to measure risk, severity, and prognosis of Alzheimer's disease and related dementias (ADRD) and Parkinson's disease (PD) using brain imaging data. Neurodegenerative disorders affect multiple body functions and exhibit significant etiology and clinical presentation variation. Patients with these conditions may experience prolonged asymptomatic periods. Disease-modifying therapies are most effective during the early asymptomatic stage of the disease, making early intervention a crucial factor. However, the lack of biomarkers for early diagnosis and disease progression monitoring remains a significant obstacle to achieving this goal. We leveraged disease specific cohorts ADNI (1,826 images from 731 dementia participants) and PPMI (684 images from 329 PD participants) to develop a machine learning classifier for AD and PD detection using T1w brain imaging data. We obtain disease-specific imaging scores from these trained models using the normalized disease probability score. In a sizeable external biobank, UK Biobank (42,835 participants), we found these scores show strong predictive power in determining the occurrence of PD or dementia during 5 year followup. The occurrence of PD increased stepwise over ascending imaging score quantiles representing heterogeneity within the PD population. Imaging scores are also associated with pathological and clinical assessment measures. Our study indicates this could be a single numeric indicator representing disease-specific abnormality in T1w brain imaging modality. The association of imaging score with the polygenic risk score of related disorders implies the genetic basis of these scores. We also identified top brain regions associated with dementia and Parkinson's disease using feature interpretation tools.

4.2.3 Implications of all the available evidence

The findings should improve our ability to create practical passive surveillance plans for individuals with a heightened risk of occurrence of neurodegenerative disease. We have shown that imaging scores complement other risk factors, such as age and polygenic risk score for early detection. The integrated model could serve as a tool for early interventions and study enrollment. Understanding the genetic basis of imaging scores can provide valuable insights into the biology of neurodegenerative disorders.

4.3 METHODS

4.3.1 Study design

We trained a supervised classification model using brain structural features extracted from T1w MRI brain images in PD specific and AD specific cohorts. Normalized disease probability scores were obtained after inference with a trained machine learning model. These scores were evaluated as a quantitative disease marker of ADRD and PD in an external large biobank.

We trained and internally validated the AD classification model using the T1w MRI data from Alzheimer's Disease Neuroimaging Initiative (ADNI, https://adni.loni.usc.edu/) cohort, which also allowed for examining the association of imaging scores with AD biomarkers and assessment tests. It consists of 731 dementia cases and 925 controls. Similarly for PD, data from Parkinson's Progression Marker Initiative (PPMI, http://www.ppmi-info.org/) was used that includes 319 PD cases, 145 controls. We assessed the association of imaging scores with relevant clinical outcomes in 42835 participants from the UK Biobank (UKB, https://www.ukbiobank.ac.uk/) with T1w brain MRI. All participants and their study partners provided their consent, accepting their engagement for the data collection. The study protocols for ADNI, PPMI and UKB were approved by the Institutional Review Board. Access and use of data from the UK Biobank was approved under application number 33601.

4.3.2 Study participants

The study included two disease specific cohorts (ADNI, PPMI) and an external large biobank (UK Biobank). ADNI enrolls participants between the ages of 55 and 90 recruited at 57 sites in the United States and Canada. After obtaining informed consent, participants undergo a series of initial tests repeated at intervals over subsequent years, including a clinical evaluation, neuropsychological tests, genetic testing, lumbar puncture, and MRI and PET scans. PPMI Clinical protocol is designed to acquire comprehensive longitudinal within-participant data from approximately 4,000 participants enrolled at about 50 sites worldwide. The quantitative marker's performance was assessed in the UK Biobank, a community-based cohort of over 500,000 individuals, mainly of British self-reported ethnicity, between the ages of 40 and 69, recruited from across the UK between 2006 and 2010. We only included participants with T1w MRI data available, randomly selected, avoiding the possibility of selection bias. The details of the number of participants used in the analysis are shown in Figure 4.1.

4.3.3 ADRD and PD diagnosis

In the PPMI cohort, all PD patients fulfilled the UK Brain Bank Criteria [60] and healthy control subjects had no clinical signs suggestive of parkinsonism, no evidence of cognitive impairment, and no first degree relative diagnosed with PD. For the ADNI cohort, the Alzheimer's disease (AD) diagnosis is based on the criteria established by the National Institute on Aging and the Alzheimer's Association (NIA-AA). The diagnosis of MCI was based on the core clinical criteria for MCI established by the NIA-AA workgroups [116] and the criteria modified from the criteria proposed by Petersen et al. [117]. Since the accurate confirmation of dementia type is done using post-mortem brain we are considering the AD cases in ADNI as ADRD rather than specific to AD. In the UK Biobank, the diagnosis date and the participants with all cause dementia were identified using Field 42018 and Field 42032 was used for PD (for detailed description of Field in UKB refer to `https://biobank.ndph.ox.ac.uk/showcase/`).

4.3.4 Clinical features obtained from T1w brain imaging data

We use the Advanced Normalization Tools (ANTs) pipeline [91] to extract structural features such as the volume and area of different brain regions from the Magnetic Resonance Imaging (MRI) T1w image. The ANTs pipeline has significantly improved performance in age and gender prediction compared to its counterpart Freesurfer [91]. Further, it can extract features from more nuanced and smaller brain regions. We used ANTsPyT1w available at `https://github.com/stnava/ANTsPyT1w` to pre-process the images. The complete pipeline yielded 869 features from each image using segmentation based on various brain atlases. After filtering and excluding 79 features who had more than 1% missing values, we trained the ML model using 790 continuous features. We performed min-max normalization to scale the range of all features between 0 and 1. Our study is the first to have a consistent pipeline for extracting such a large number of features from two large neurodegeneration specific and biobank scale datasets.

4.3.5 Development of the machine learning model

We develop a stacked ensemble of three supervised machine learning algorithms (Neural network, XGBoost, Logistic regression) to train a classification model for Parkinson's disease (PD) and Alzheimer's disease (AD) in disease specific cohorts. To train an ensemble model, we followed the strategy proposed in AutoML tool h2o. The best performing model (based on performance metric) was selected after extensive hyperparameter tuning to build the final ensemble model. Feature selection was performed using Least Absolute Shrinkage and Selection Operator (LASSO) [118] as a part of the hyperparameter tuning procedure. The effectiveness of our classifier was validated using a 3-fold cross validation approach. Model performance was evaluated on the basis of various metrics, including accuracy, area under the receiver operating characteristic curve (AUC), sensitivity and specificity. Although longitudinal data was used for training the model, we ensured that the cross-validation folds were split in such a way that the same individual did not appear in both the training and test folds. This approach ensured that the model performance reports were free from bias. The 95% confidence intervals (CIs) on diagnostic and prognostic performance estimates were calculated by use of a percentile bootstrap with 10,000 samples. We employed the Shapley Additive Explanations (SHAP) [119] (refer to chapter 2, section 2.3.10 for more details) approach to assess the impact of each feature on the machine learning model predictions. SHAP values are derived from game theory and provide an approximation of a feature's effect on the model. SHAP enhances understanding by creating accurate explanations for each observation. The interactive website (`https://ndds-brainimaging-ml.streamlit.app/`) was developed as an open-access and cloud-based platform for researchers to investigate the top features of the machine learning models developed and how these may influence the disease probability scores. Finally, the probability scores were normalized using log-odds transformation to generate the disease specific quantitative markers.

4.3.6 Features from UK Biobank cohort

We computed ADRD and PD imaging scores for all participants in UK Biobank who have available T1w brain MRI. In our analysis, we also considered known risk factors associated with neurodegenerative disorders comprising age at recruitment (Field 21022), Date of attending assessment center (Field 53), Sex (Field 31), Townsend deprivation index at recruitment (Field 22189), Standard polygenic risk score (PRS) for Alzheimer's disease (Field 26206) and Standard PRS for Parkinson's disease (Field 26260) available in UKB database.

4.3.7 Clinical assessments and MCI convertors to Dementia

We evaluated the association of imaging scores with standard clinical assessments used for diagnosing dementia, such as MoCA score, MMSE, and ADAS-Cog, available in the ADNI database. Additionally, we utilized pathological biomarkers of AD, including AV45 and FDG, to validate the imaging scores. We also assessed the predictive value of imaging scores for the conversion from MCI to dementia. To validate the scores, we extracted MDS-UPDRS-III, MoCA, and SerumNfl biomarkers from the PPMI database.

4.3.8 Statistical analysis

We employed a Cox proportional hazards regression model to examine the correlation between imaging scores and the clinical diagnosis time of participants who had imaging data before diagnosis. These models were adjusted for relevant covariates. To estimate survival curves, we used imaging score quantiles. We reported the hazard ratio along with the corresponding p-values and 95% CI. The overall performance was measured using the concordance index, while the time-dependent AUC metric, using sksurv, was used to evaluate the model's performance in terms of stratification based on time from the event. Linear regression was used to associate the polygenic risk score with the imaging score, and Pearson correlation coefficient was used to determine the correlation between imaging scores and disease monitoring clinical assessment tests. All statistical tests and plots were created using Python (version 3.8). Figure 4.1 highlights the workflow for our analysis.

4.4 RESULTS

Feature extraction using AntsPyT1w from T1w MRI images was completed for 2,546 participants (22,004 images) in the ADNI dataset, 1,316 participants (2,601 images) in the

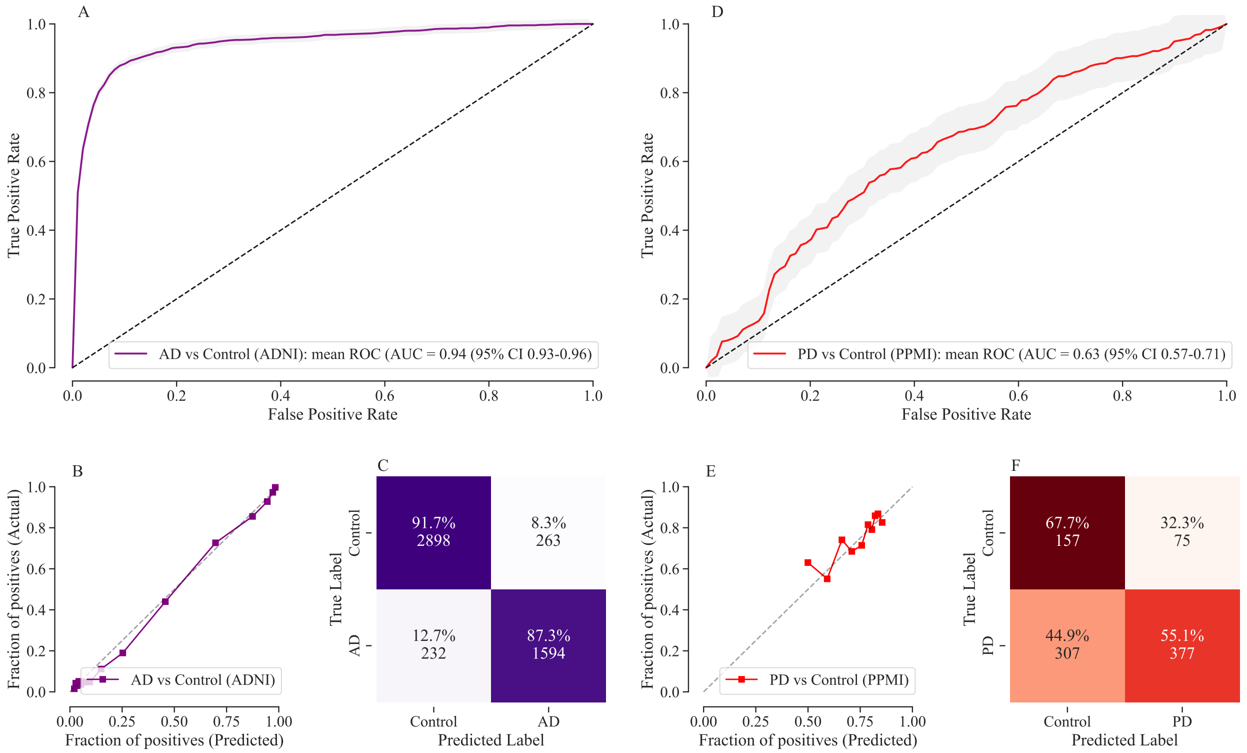


Figure 4.2: **Performance of the machine learning model for the detection of AD in the ADNI and PD in the PPMI cohort following stratified 3-fold CV (based on individuals) evaluation scheme.**(A), (D) The machine learning model discriminated Dementia from healthy controls with cross-validated AUROCs of 0·94 (95% CI 0·93–0·95), and PD from healthy controls with AUROCs of 0.63 (95% CI 0·58–0·69). (B), (E) Shows the calibration curves for trained classifiers, demonstrating that the model is well calibrated and not making overconfident or underconfident probabilistic predictions. (C), (F) Shows the confusion matrix on cross-validation folds for AD and PD detection respectively at the threshold with maximum F1 score. CV, cross-validation; AD, Alzheimer's Disease; PD, Parkinson's Disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; PPMI, Parkinson's Progression Markers Initiative; AUROC, area under the receiver operating characteristic curve

PPMI cohort, and 42,925 participants (46,124 images) available in the UK Biobank cohort. Each image has an automated quality score from the ANTs pipeline that is used to remove noisy images. There may be multiple images for a single participant on a particular visit, so we chose the best quality images out of these. Further, we excluded participants who were less than 40 years old. We trained an ADRD classification model on 1,826 dementia and 3,161 control image samples from the ADNI cohort. We also trained a PD classification model on 684 dementia and 232 control image samples from the PPMI cohort (demographics shown in Table 4.1).

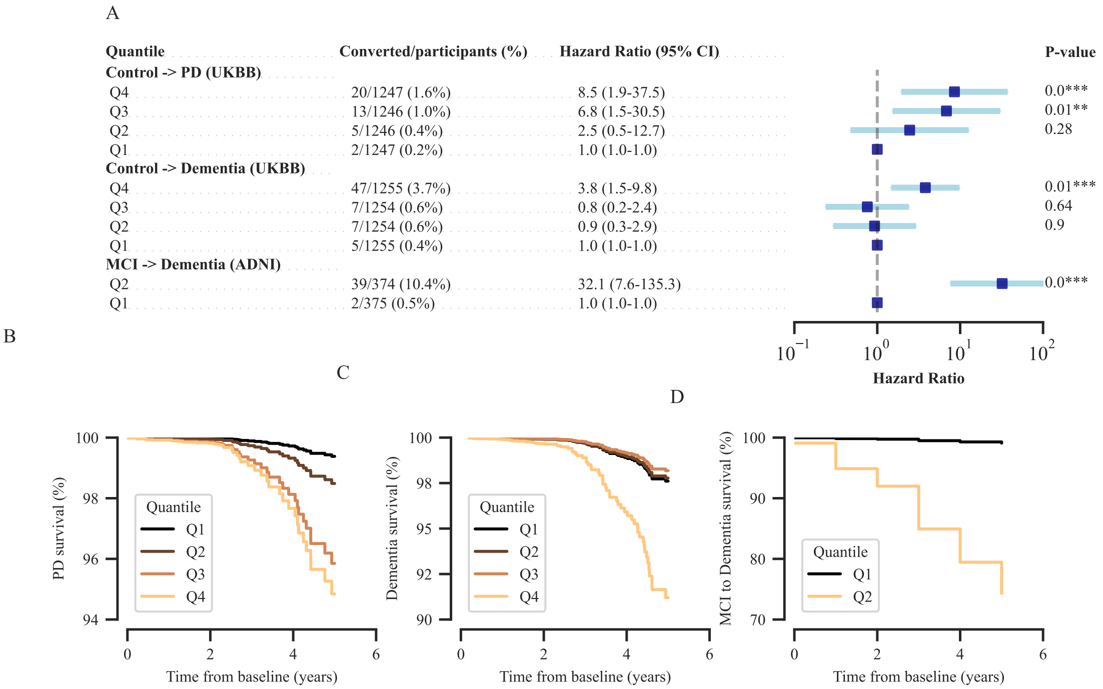


Figure 4.3: **Testing utility of AD/PD imaging scores as a risk, prognostic, and monitoring biomarker.** Imaging scores were stratified into multiple quantiles and adjusted HRs were compared with the lowest quantile. (A) Plot summarized the conversion from healthy control to PD (PD risk), healthy control to Dementia (Dementia risk) and MCI to Dementia (prognosis) with adjusted HR using Cox PH survival model. Both HR and converted percentage increased over ascending imaging score quantiles. The effect size is significantly different between the highest and lowest quantile for all the conversion as depicted in the forest plot. (B), (C), and (D) Covariate-adjusted survival curves for patients for different quantiles with duration (in years) from the conversion on the x-axis and fraction of individuals surviving free of conversion on the y-axis. Higher imaging score quantiles had lower survival for conversion as compared with lower quantiles. HR, hazard ratio; Cox PH, Cox Proportional-Hazards;

We correctly distinguish images with ADRD based on imaging features with an AUC of 0.94 (95% CI: 0.93-0.96) and PD with an AUC of 0.63 (0.56-0.70) at cross-validation (subject based) (Figure 4.2A, Figure 4.2D). Both the models are very well calibrated for ADRD detection model and PD (Figure 4.2B, Figure 4.2E, Table 4.2). The well-calibrated model is necessary in our study as we focus on the usefulness of probabilistic scores obtained from machine learning models. The quantification scores allow us to extend the model trained on post-diagnosis phase to develop a model for early detection of AD and PD during pre-diagnostic phase.

We evaluated the association between ADRD/PD imaging score and the occurrence of

Table 4.1: Characteristics of participants in the training and internal validation set

	PPMI		ADNI	
	Controls (I=232, N=145)	**PD (I=684, N=319)**	**Controls (I=3,161, N=925)**	**Dementia (I=1,826, N=731)**
Mean age at baseline (SD), years	62 (10)	62 (9)	73 (7)	76 (8)
Mean age at T1w MRI performed (SD), years	62 (10)	63 (9)	75 (7)	76 (8)
Sex				
Female	47 (32%)	111 (35%)	457 (56%)	237 (42%)
Male	98 (68%)	208 (65%)	358 (44%)	334 (58%)
Race				
White	136 (93%)	296 (93%)	803 (87%)	683 (93%)
Black or African American	7 (5%)	5 (2%)	75 (8%)	26 (4%)
Asian	1 (1%)	6 (2%)	26 (3%)	16 (2%)
Other / unknown	2 (1%)	12 (3%)	24 (2%)	6 (1%)
Baseline clinical outcomes				
MOCA	28 (1)	27 (2)	26 (3)	17 (5)
MMSE	NA	NA	29 (1)	23 (3)
MDS-UPDRS III	1 (2)	22 (9)	NA	NA
ADAS-Cog-13	NA	NA	9 (4)	29 (8)

Table 4.2: Classification performance of machine learning model for PD and AD detection

	AUROC	**Sensitivity**	**Specificity**	**Accuracy**	**Balanced Accuracy**	**PPV**	**NPV**
PD detection (3-fold CV)	0.63 [0.58-0.69]	0.55 [0.51-0.59]	0.68 [0.62-0.73]	0.58 [0.55-0.61]	0.61 [0.58-0.65]	0.83 [0.80-0.87]	0.34 [0.30-0.38]
AD detection (3-fold CV)	0.94 [0.93-0.95]	0.87 [0.86-0.89]	0.92 [0.91-0.93]	0.90 [0.89-0.91]	0.89 [0.89-0.90]	0.86 [0.84-0.87]	0.93 [0.92-0.93]

all cause dementia and PD in UK Biobank during the pre-diagnostic phase using survival analysis. In UKB, out of 92 dementia patients with imaging data, 76 have images collected during the pre-diagnostic phase and 16 have during post-diagnosis phase. For PD, 46 have image collection during pre-diagnostic and 50 during post-diagnosis phase. ADRD imaging score was associated with a quantitative increase in risk of getting dementia, with an adjusted HR of 1·5 (95% CI 1·4–1·6; $p < 0.0001$) per SD increase in ADRD imaging score. Also, PD imaging score was associated with a quantitative increase in risk of getting PD diagnosis, with an adjusted HR of 1·5 (95% CI 1·4–1·6; $p < 0.0001$) per SD increase in PD imaging score. Prevalence and risk of getting clinical diagnosis increased stepwise over ascending imaging quartiles: two (0.2%) of 1247 participants who had follow up in the bottom quartile (Q1), and 23 (1.9%) of 1247 in the top quartile (Q4; 56 [20–158]; $p < 0.0001$; Figure 4.3). We observed a phenotypic continuous spectrum for PD in which the imaging score (abnormality in T1w image) varies across the PD population during the pre-diagnostic phase. For dementia, we observed more homogeneous patterns, as the majority of the subjects (in quartile 1) demonstrated abnormality in T1w during the pre-diagnosis stage.

We used time-dependent ROC and concordance index to characterize the discrimination potential of the imaging scores (Table 4.3). We observed that the predictive importance of imaging score in determining survival is higher for subjects who converted within three years from baseline as compared to those who converted after three years. It is consistent

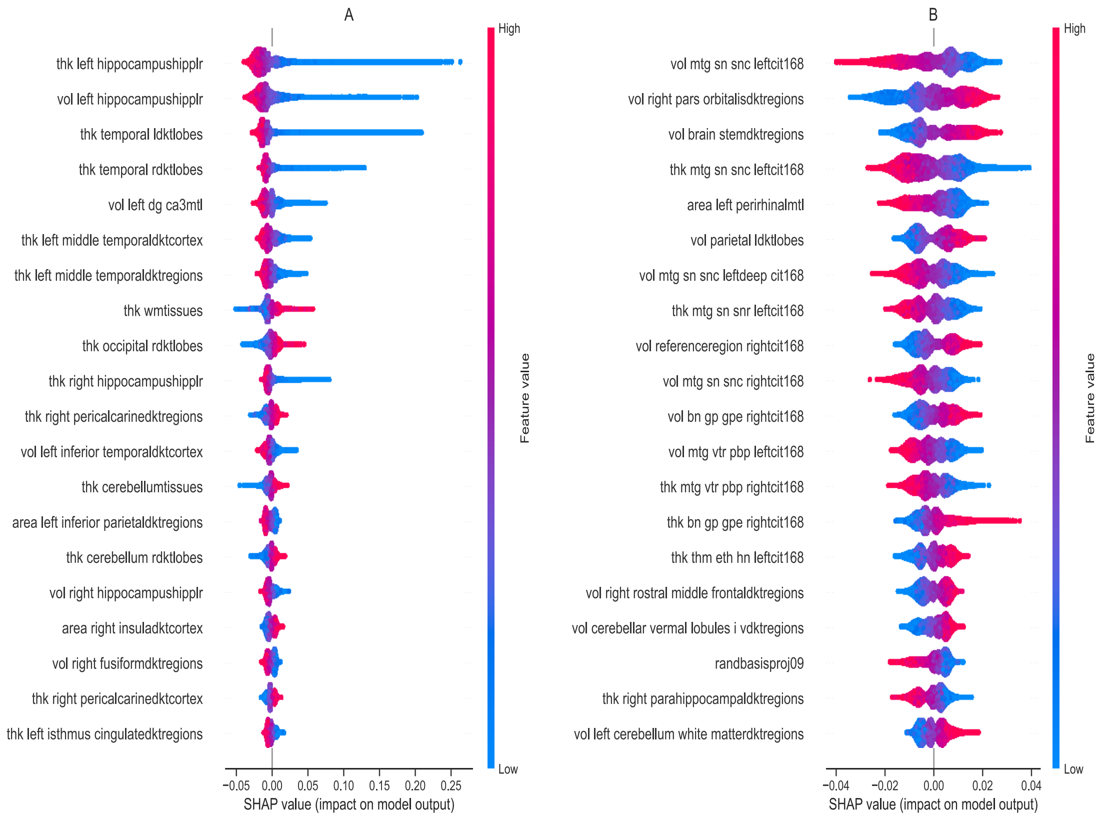


Figure 4.4: **SHAP Interpretation of imaging scores.** (left plot) Distribution of the top-20 features that had the most substantial effect on the ADRD imaging score (disease probabilities). Each point represents a patient and the amount of effect on model output for each feature depends on its SHAP value. For example, the effect of the hippocampus cortical thickness feature on model output is large and positive (high risk) when the patient has low values for hippocampus cortical thickness (more blue points are on the right side). Similarly, right plot shows the top features for PD imaging scores.

Table 4.3: Survival analysis detailed statistics for risk/prognosis prediction

	Imaging HR (95% CI)	Imaging P-value	Concordance Index (95% CI)	Mean t-AUC (95% CI)	Mean t-AUC (95% CI) <3 years	Mean t-AUC (95% CI) >= 3 years
PD risk prediction						
A + S + T + PD_PRS	NA	NA	0.83 [0.78-0.89]	0.87 [0.82-0.92]	0.84 [0.77-0.93]	0.91 [0.86-0.97]
PD_Imaging + A + S + T + PD_PRS	2.33 [1.55-3.50]	4.8e-5	0.86 [0.81-0.92]	0.89 [0.85-0.94]	0.87 [0.81-0.96]	0.92 [0.89-0.99]
AD risk prediction						
A + S + T + AD_PRS	NA	NA	0.82 [0.76-0.89]	0.84 [0.77-0.90]	0.78 [0.66-0.92]	0.89 [0.82-1.01]
ADRD_Imaging + A + S + T + AD_PRS	1.76 [1.50-2.05]	1.6e-12	0.86 [0.79-0.93]	0.86 [0.80-0.92]	0.81 [0.67-0.97]	0.90 [0.84-1.01]
MCI =>Dementia risk prediction						
A + S + BV + AD_PRS	NA	NA	0.52 [0.41-0.61]	0.51 [0.37-0.63]	0.55 [0.30-0.86]	0.50 [0.33-0.64]
ADRD_Imaging + A + S + BV + AD_PRS	2.99 [2.31-3.86]	5.7e-17	0.73 [0.64-0.84]	0.74 [0.63-0.87]	0.65 [0.40-0.97]	0.73 [0.60-0.85]

for both dementia and PD conversion. It is an important result as it guides the healthcare system to determine the duration between image collections. The contribution of imaging

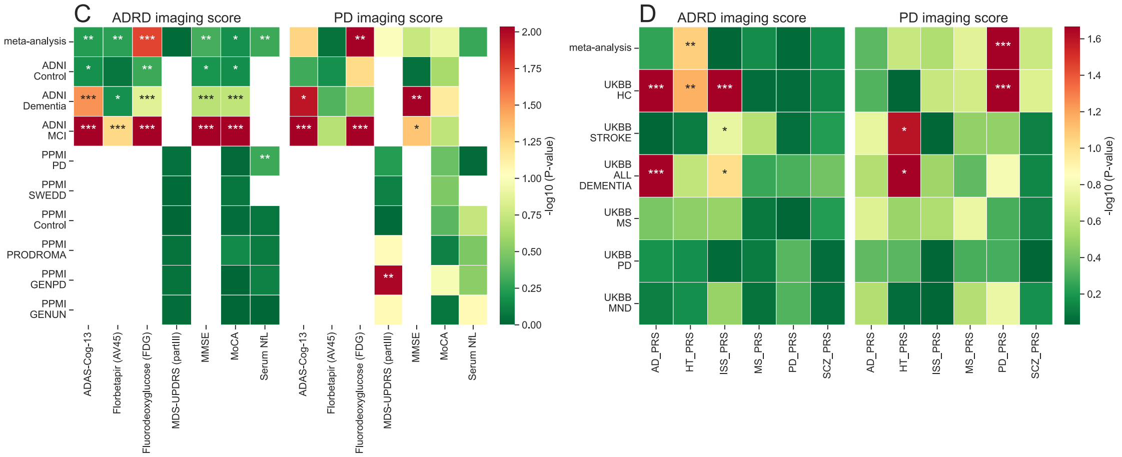


Figure 4.5: **Association testing of AD/PD imaging scores with pathological biomarkers and genetic risk scores.** (left plot) Shows the heatmap plot of P-values for the association testing between ADRD/PD imaging scores with clinical outcomes in ADNI and PPMI cohort stratified based on diagnosis status. (right plot) Heatmap plot showing association test of imaging scores with polygenic risk scores of related diseases in the UK biobank cohort. ADAS−Cog−13, Alzheimer's Disease Assessment Scale−Cognitive Subscale; MoCA, Montreal Cognitive Assessment; MMSE, Mini-Mental State Examination; MDS-UPDRS III, MDS-Unified Parkinson's Disease Rating Scale Part III; Serum Nfl, Serum neurofilament light; PPMI-SWEDD, Patients with scans without evidence of dopaminergic deficit in PPMI; PPMI-PRODROMAL, prodromal PD patients; PPMI-GENPD, Genetic affected PD; PPMI-GENUN, Genetic unaffected PD; AD_PRS, polygenic risk score of Alzheimer's disease; HT_PRS, polygenic risk score of Hypertension; ISS_PRS, polygenic risk score of Ischemic stroke; MS_PRS, polygenic risk score of Multiple Sclerosis; PD_PRS, polygenic risk score of Parkinson's disease; SCZ_PRS, polygenic risk score of Schizophrenia; UKBB-HC, healthy control in UK biobank (with no NDDs); UKBB-STROKE, Stroke patients in UK biobank; UKBB-ALL_DEMENTIA, all cause dementia patients in UK biobank; UKBB-MS, Multiple Sclerosis patients in UK biobank; UKBB-PD, Parkinson's disease patients in UK biobank; UKBB-MND, Motor neurone disease patients in UK biobank; ***, P<0.001; **, P<0.005; *, P<0.01;

score is further useful in early detection of prognosis from MCI to dementia with the hazard ratio (32.1 [7.6–135.3]; $p < 0.0001$) for quartile 1.

SHAP values provide in depth analysis of machine learning classifiers by highlighting the top discriminating features of ADRD and PD. The biological plausibility of the models is supported by the feature importance plots, which highlight regions known to be earliest regions affected during AD and PD. In ADRD, features related to the hippocampus and temporal brain regions have the greatest discriminatory power (Figure 4.4A). Midbrain regions, including substantia nigra, brain stem, and parietal brain regions, help distinguish PD from healthy people (Figure 4.4B).

To assess the potential of imaging scores as a monitoring tool during the post-diagnosis phase, we conducted association testing with clinical and pathological biomarkers in the PPMI and ADNI cohorts. We found a strong association between the imaging scores and cognitive measures (i.e, Alzheimer's Disease Assessment Scale−Cognitive Subscale, Montreal Cognitive Assessment tests), especially for the AD imaging score, and a weaker association for the PD imaging score (Figure 4.5C). In addition, we observed a strong association between the imaging scores and the MoCA ($P<0.001$), MMSE ($P<0.001$), and ADAS-Cog-13 ($P<0.001$) scores, while the correlation with the MDS-UPDRS- III score was not as strong for the PD imaging score (Figure 4.5C) ($P>0.05$). The imaging phenotypes can further help us assess the genetic architecture of neurodegenerative disorders. We found a significant association between imaging scores and disease-specific polygenic risk scores obtained from case-control genome wide association studies (GWAS) in the UK Biobank cohort (Figure 4.5D). The p-value for the association testing of ADRD imaging score and AD polygenic risk score, as well as PD imaging score and PD polygenic risk scores, was found to be less than 0.001.

4.5 DISCUSSION

This study uses the brain imaging data and machine learning approach to generate disease specific scores. We have shown the potential of imaging scores in pre-diagnostic prediction, addressing heterogeneity and monitoring disease severity. We observed that not all PD or all cause dementia subjects show abnormality in brain anatomical features during their prediagnostic phases. Finally, the association with clinical and pathological biomarkers shows that these scores can also be used to monitor disease progression. This study presents a general framework to generate machine learning based disease scores from complex data modalities such as imaging and is easily extendable to other disorders.

Identifying individuals during the asymptomatic phase is the key for successful intervention using disease modifying treatments. Since, the brain atrophy starts much before the symptomatic phase of a disease therefore structured brain imaging may help. As we observed significant variability in terms of imaging; this might reflect some differences in underlying pathologies. Therefore, we believe that these scores have the potential of intervention during study enrollment to have more homogeneous patients going into the trial. Imaging scores can work as complementary measures in addition to pathological markers such as α-synuclein seed amplification [120], to select the patients in precision medicine trials. Our models can be used as potential therapeutic engagement readouts. However, it is important to note that a single measure or modality may not be sufficient to capture all the symptoms of such disorders. Therefore, using multiple modalities (such as combining imaging, behavioral, and clinical data) could provide a more comprehensive and useful approach to assessing therapeutic engagement. Passive surveillance is another crucial aspect that has to be incorporated in health care systems [121, 122]. It involves monitoring an individual's data over time to detect any abnormalities that may occur, even in the absence of symptoms. With the cost effectiveness and availability of T1w brain MRI imaging, our models have the potential to track patients' brain health more effectively. Ultimately, this could result in improved patient outcomes and more efficient use of healthcare resources.

In our current analysis, we also showed association of disease imaging scores with polygenic risk scores of related disorders. There could be direct implications of performing GWAS on imaging in a UK biobank. First, we can understand the overlaps across neurodegenerative disorders at the gene level. Second, we might find novel risk loci using imaging scores as we can have more accurate phenotype, reducing proxy cases which is common in current GWAS studies. ML based probabilities are useful in identifying new genetic loci for glaucoma and chronic obstructive pulmonary disease [123]. Transparency and reproducibility is a critical aspect of science [124]. It becomes further more important for machine learning models due to the dependency on data and black box nature of machine learning models [125, 126, 127]. To facilitate this, we have developed an interactive website that allows researchers to explore the top features contributing to predicting disease probabilities. Further, we incorporated a model perturbation analysis feature in our website where researchers can manually play with the features and observe how the prediction changes. Therefore, the transparency of our approach, using SHAP values, and the contributions of data types move the research community away from black-box predictors.

Although this study presents the potential of machine learning based biomarkers on complex data such as imaging, much remains to be established. First, the limited availability of data from non-European participants' dataset can introduce inherent bias. Second, the diagnosis that we have used both for validation and model training are not pathologically confirmed. We put efforts to avoid overfitting in our models so that the ML model won't fit to misdiagnosed cases. However, pathology confirmed cases are particularly important for the validation purpose. Third, in this work, we are considering ADRD as a single group but it is very important to have biomarkers that can differentiate between dementia types. Fourth, the analysis of common brain regions affected in both Alzheimer's and Parkinson's disease would be useful future work. Finally, co-pathology is very common across neurodegenerative disorders [128, 129] which should be considered in future studies.

In summary this work presents an approach to generate disease biomarkers using machine learning and brain imaging data. This study is a step forward toward utilizing sophisticated machine-learning paradigms to facilitate the early disease detection, prognosis and monitoring of disease progression.

So overall, we have developed various machine learning models to address the challenges presented by neurodegenerative disorders. However, machine learning tools also present challenges, with the topmost concerns being transparency, reproducibility, and accessibility of the models.

CHAPTER 5: TRANSPARENT, REPRODUCIBLE AND ACCESSIBLE SCIENCE

In this chapter, we will discuss the measures we have taken to make all our machine learning models more accessible. We believe that this is a necessary task if we really want to bring machine learning into medical practice.

5.1 INTRODUCTION

Transparency, reproducibility and accessibility is a critical aspect of science. It becomes further more important for Artificial Intelligence due to the dependency on data and black box nature of machine learning models [125, 126, 127]. We are facing a major reproducibility crisis in science [124], with only 6% of the papers in the Artificial Intelligence community sharing the algorithm's code [130]. It is particularly crucial in medical science because of the expense of drug discovery [131] and from an ethical point of view [132, 133]. Therefore, we performed code sharing with proper documentation and the public sharing of research results/tools through web applications. In all our works, we used open-source development tools to release our code and research results and make it accessible to research and medical professionals. The idea to keep up with open-science philosophy of transparent and accessible knowledge to everyone.

5.2 METHODS

In the following section, we discuss the tools we used to make our research transparent, reproducible and accessible. The interactive websites were developed as an open-access and cloud-based platform for researchers to investigate the research output.

5.2.1 Transparency

Transparency of machine learning models involves explaining how machine learning systems make decisions or predictions, which can be challenging due to the high level of complexity in machine learning models. Merely showing accuracy estimates of these models can hide their complexity. It involves making the algorithms and models used in AI systems open and accessible, as well as disclosing the data sets used to train the models. Providing transparency in ML can help medical professionals to better understand how the ML system arrived at a particular decision or prediction, which can build trust in the technology. As machine learning algorithms increase in size, incorporating a larger number of parameters, with the use of neural networks, ensemble classifiers, and other complex methods, transparency becomes increasingly important.

In this dissertation, we employ both supervised and unsupervised machine learning algorithms. However, ensuring transparency for both types of algorithms requires different approaches and tools. In this section, we describe how we address some of the challenges posed by these models from the viewpoint of transparency. When it comes to evaluating unsupervised machine learning algorithms, visual inspection is typically the main approach. To facilitate this process, we have developed a platform that allows for easy inspection of the results. Our platform includes a color-coding system that helps to guide evaluation of the lower-dimensional space based on available metadata. Given that metadata can often contain a large number of features, and researchers may want to understand the underlying patterns that are driving the observed results, it is crucial to provide sophisticated tools that simplify this process. The visualization of lower-dimensional data space can guide us in various tasks, such as identifying clusters, sub-structures, and outliers, detecting batch effects, and quality control measures to perform reliable and accurate downstream analyses [134, 135, 136].

For supervised models that have been trained, interpreting the features and analyzing model perturbations are important aspects for improving transparency. We've created a dashboard that enables users to perform a “what-if” analysis on the features used to train the model. Shapley additive explanations (SHAP) [119] approach was used to evaluate each feature's influence in the ensemble learning. This approach, used in game theory, assigned an importance (Shapley) value to each feature to determine a player's contribution to success. Shapley explanations enhance understanding by creating accurate explanations for each observation in a dataset. They bolster trust when the critical variables for specific records conform to human domain knowledge and reasonable expectations [137, 138]. We used the one-vs-rest technique for multi-class classification. Based on that, we trained a separate binary classification model for each class. By examining the top features of the model developed in this study, users can gain insight into how these features may affect the classification (or sometimes, misclassification) of a given sample.

5.2.2 Reproduciblility

Small errors or missing details in the computer code could have significant impacts on the machine learning training and evaluation of results. Therefore, it's important to be transparent about the actual computer code used to train a model, the process of hyper-parameter tuning and the final optimized parameters, to ensure reproducible research. To facilitate replication and expansion of our work, we have made the notebook publicly available on GitHub. It includes all code, figures, models, and supplements for this study. The code is part of the supplemental information; it includes the rendered Jupyter notebook with full step-by-step data preprocessing, statistical, and machine learning analysis.

Computational tools and code availability Most of the data-analysis work was done in Python (version 3.8) using open-source libraries (NumPy [version 1.20.3], pandas [version 1.2.5], matplotlib [version 3.4.2], seaborn [version 0.11.1], plotly [version 4.14.2], scikit-learn [version 0.24.2], UMAP [version 0.5.0], XGBoost [version 1.4.2], LightGBM [version 3.2.1], GenoML [version 2v1.0.0b11], and TensorFlow [version 2.4.0]). For all our work, we made our code publicly available on GitHub to facilitate replication and future expansion of our work.

5.2.3 Accessiblity

Web dashboard Streamlit is an open-source Python library that makes it easy to create and share, custom web applications for machine learning and data science (`https://streamlit.io/`). It is useful in deploying data application with interactive facility. It is a commonly used tool to build prototypes of data science projects. Further, it allows sharing open-access dashboard using cloud-based platform making it easily accessible. All the project developed as part of this dissertation includes such an application dashboard that allows to inspect the models and data used for the projects.

5.3 RESULTS

In this section, we describe the list of things that we did to make our works more transparent, reproducible and interpretation. Particularly, we will discuss the features of dashboard designed to allow exploration of both unsupervised and supervised models. The web links are listed below:

- Subtypes for Amyotrophic Lateral Sclerosis (ALS): `http://bitly.ws/u5h7`
- Subtypes for Alzheimer's disease (AD): `http://bitly.ws/u5h3`
- Subtypes for Parkinson's disease (PD): `https://shorturl.at/fkvw3`
- Longitudinal Aligned-UMAP: `https://rb.gy/zf2xu`

5.3.1 Unsupervised models

The subtype identification process requires creating a multi dimensional space that captures the disease's features, for instance, the progression rate of Parkinson's disease (PD) subtypes (i.e., velocity). The data dimensional reduction methods on the complex clinical features created meaningful spatial representation of each patient's status at this time point.

Figure 5.1 displays the lower dimensional space generated by Non-negative matrix factorization approach on ADNI dataset. The dashboard enables users to choose the time point for analyzing the lower dimensional space and color the scatter plot using single measures at any specific point in time, such as ADAS-Total-Scores at baseline, as shown in Figure 5.1. For PD subtyping, we demonstrate how the latent space varies as we incorporate more longitudinal data in our approach (Figure 5.2). The dashboard developed for Aligned-UMAP work incorporates interactive dimensional reduction plots for all seven longitudinal datasets included in the analysis, as explained in detail in Chapter 3. To observe the effect of different hyper-parameter settings, we added an additional feature to Longitudinal Aligned-UMAP. Figure 5.3 depicts the significant decrease in volatility as we increase emphasis on longitudinal regularization terms.

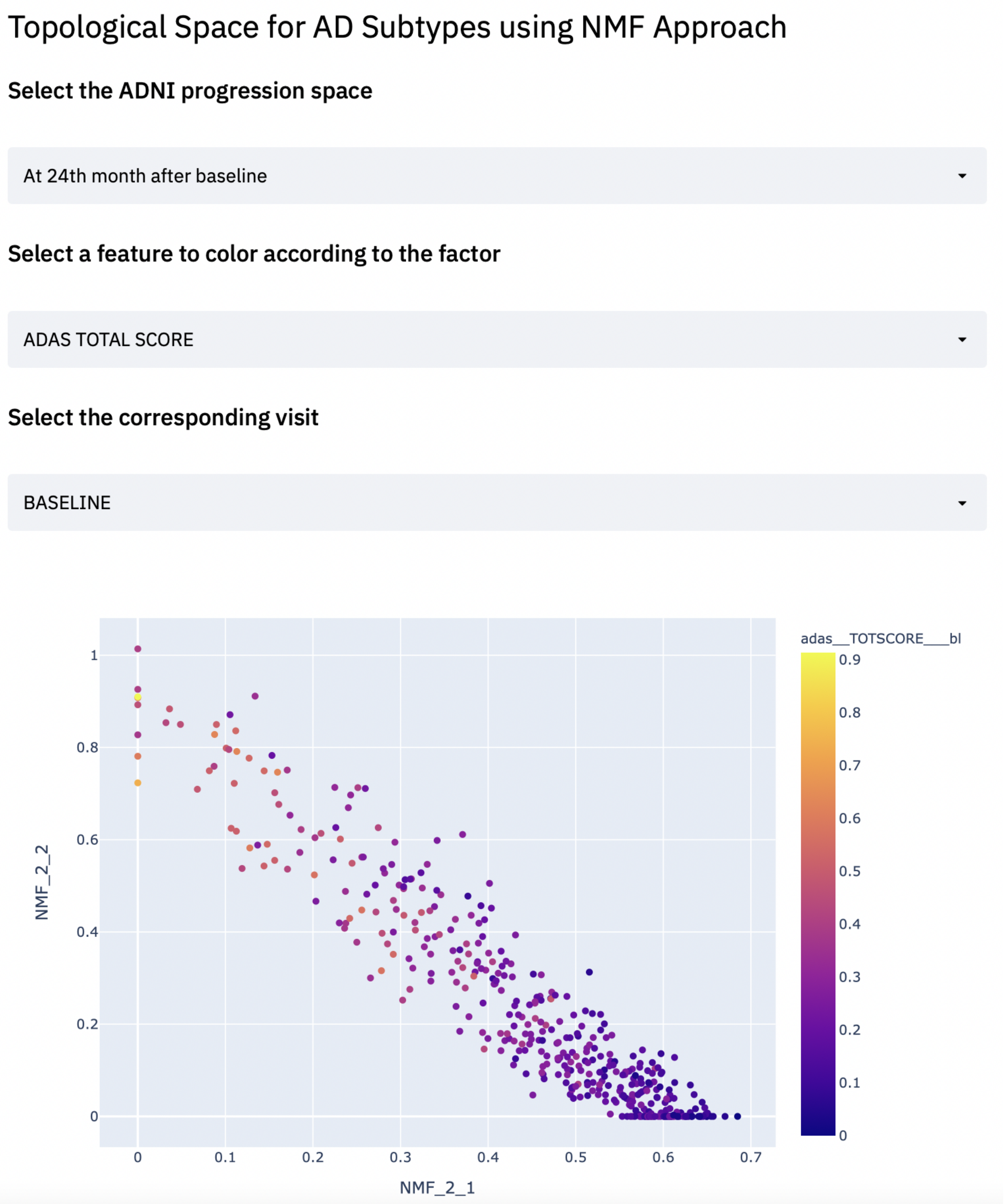
5.3.2 Supervised models

SHAP values A surrogate xgboost model [80] was trained in 70% of the data and later tested in the 30% of withheld data to evaluate the model's contributing features. The interactive website was developed as an open-access and cloud-based platform for researchers to investigate the top features of the model developed in this study and how these may influence the classification.

Figure 5.4 shows the top features involved in accurate diagnosis of PD using multi-omic dataset. It is very difficult to explain the effect of each individual feature when we have very high dimensional data for prediction. The utility of our designed tools grows significantly in such cases. Similar interactive data-driven web applications were developed for PD subtype predictions (Figure 5.5).

What-if analysis The web applications will also provide a useful result interpretation and "what-if analysis" tool, a very useful tool to investigate effects of possible factor changes or early interventions. In our web applications toolbox, we also provide Shapely values to increase the interpretability of the results.

Figure 5.6 shows the interface to provide top-20 features required for the classification model to predict subtype probabilities. The output obtained contain two items (1) the



Topological Space for AD Subtypes using NMF Approach

Select the ADNI progression space

At 24th month after baseline ▾

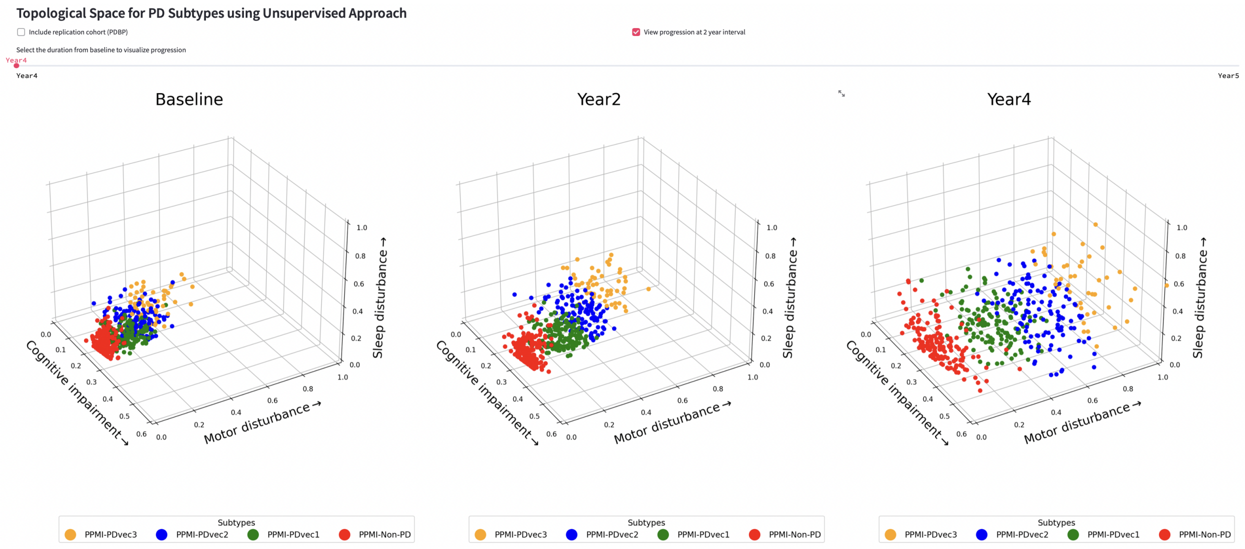
Select a feature to color according to the factor

ADAS TOTAL SCORE ▾

Select the corresponding visit

BASELINE ▾

Figure 5.1: Shows the dashboard with views of AD progression space generated based on user provided instructions. The shown scatter plot is generated using three longitudinal visits over 24 months. Coloring is done based on ADAS-total-score collected at baseline.



Topological Space for PD Subtypes using Unsupervised Approach

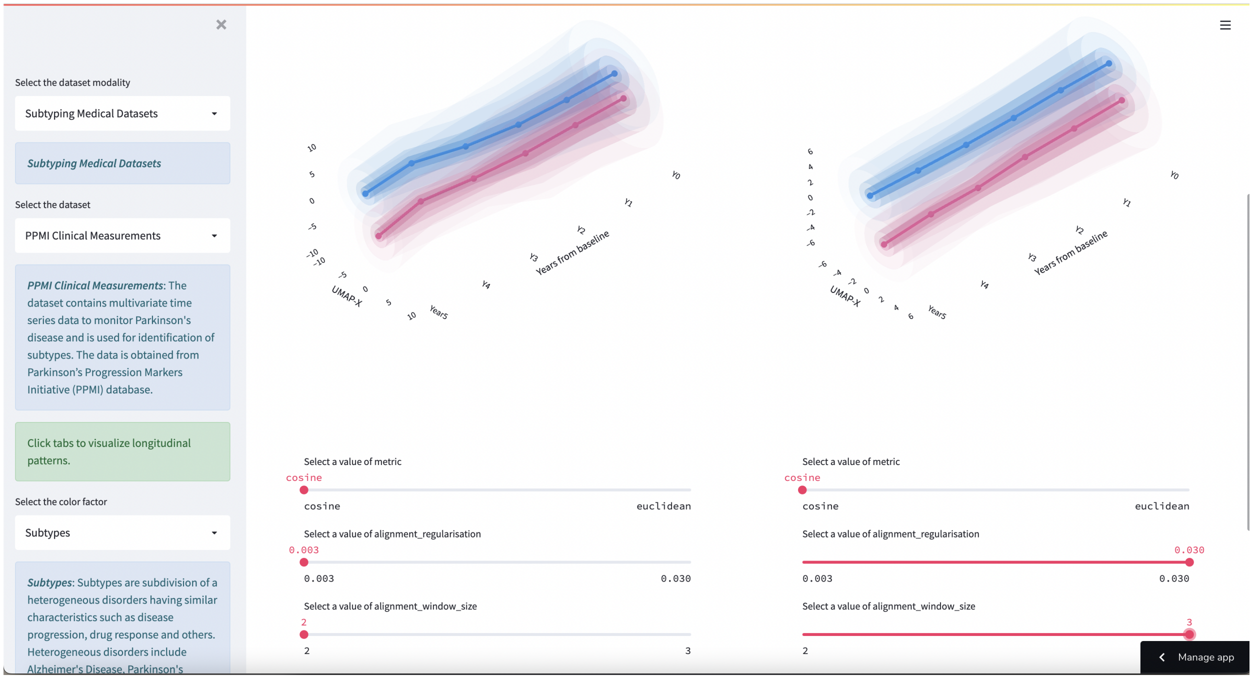
☐ Include replication cohort (PDBP)

☑ View progression at 2 year interval

Select the duration from baseline to visualize progression

Year4

Figure 5.2: Two-dimensional progression space colored by PD subtypes (observed using 5 years of longitudinal data) showing their normalized trajectory.



Select the dataset modality

Subtyping Medical Datasets

Subtyping Medical Datasets

Select the dataset

PPMI Clinical Measurements

PPMI Clinical Measurements: The dataset contains multivariate time series data to monitor Parkinson's disease and is used for identification of subtypes. The data is obtained from Parkinson's Progression Markers Initiative (PPMI) database.

Click tabs to visualize longitudinal patterns.

Select the color factor

Subtypes

Subtypes: Subtypes are subdivision of a heterogeneous disorders having similar characteristics such as disease progression, drug response and others. Heterogeneous disorders include Alzheimer's Disease, Parkinson's

Select a value of metric: cosine (cosine – euclidean)

Select a value of alignment_regularisation: 0.003 (0.003 – 0.030)

Select a value of alignment_window_size: 2 (2 – 3)

Select a value of metric: cosine (cosine – euclidean)

Select a value of alignment_regularisation: 0.030 (0.003 – 0.030)

Select a value of alignment_window_size: 3 (2 – 3)

Figure 5.3: Parameter tuning dashboard of longitudinal dimensional-reduction algorithm (Aligned-UMAP). Figure shown here is generated using user-defined input parameters of Aligned-umap on PD subtype data (i.e. alignment regularization terms).

changes in class prediction probabilities and (2) the explanation plots (force plot, decision plot (bottom) illustrating the influence of each feature on the model's prediction for a single individual (Figure 5.7). Similar dashboards were developed to predict ALS subtypes

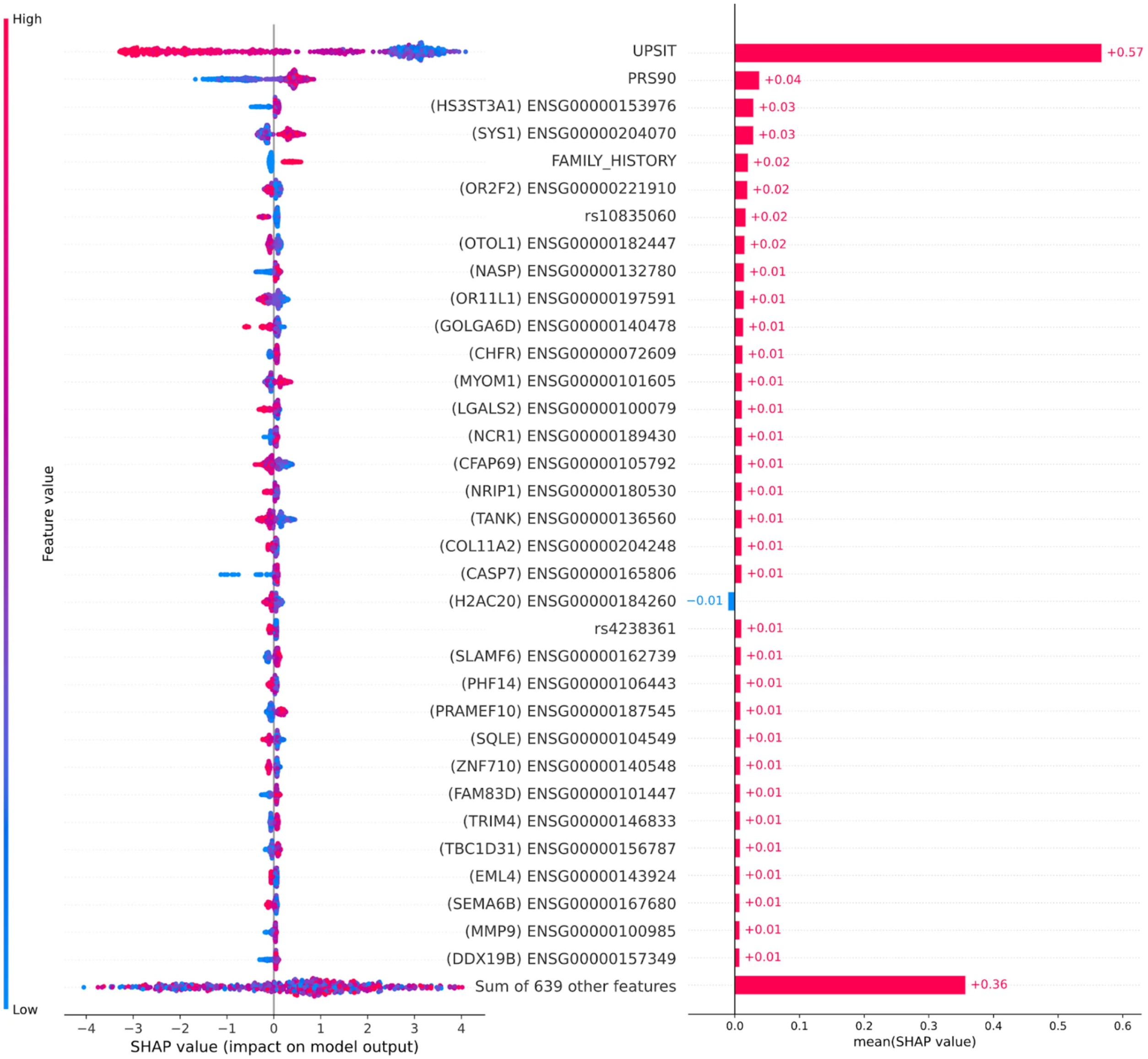


Figure 5.4: Feature importance plots for top 5% of features in data. The plot on the left has lower values indicated by the color blue, while higher values are indicated in red compared to the baseline risk estimate. Plot on the right indicates directionality, with features predicting for cases indicated in red, while features better-predicting controls are indicated in blue. SHAP Shapley values, UPSIT University of Pennsylvania smell identification test, PRS polygenic risk score.

(Figure 5.8).

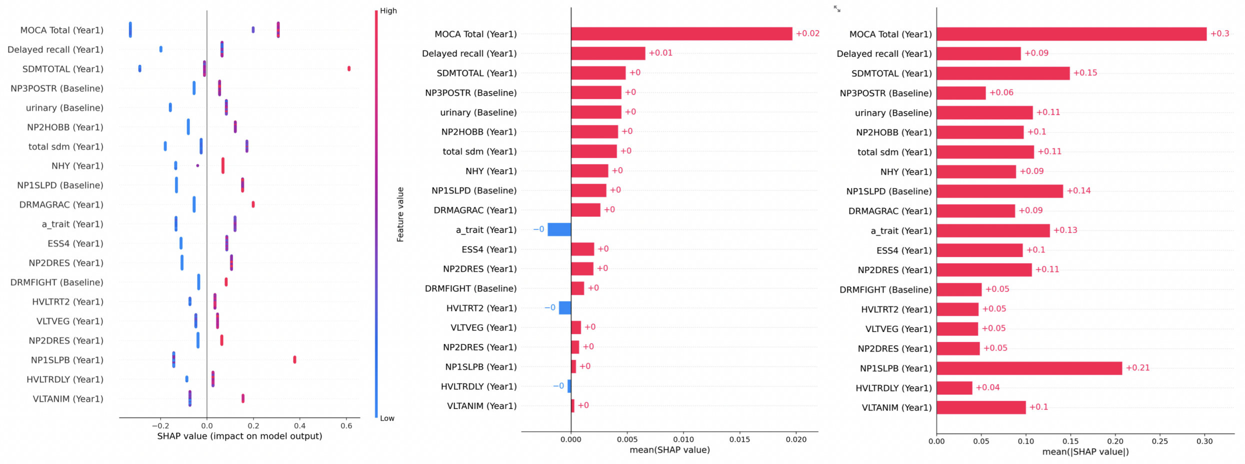
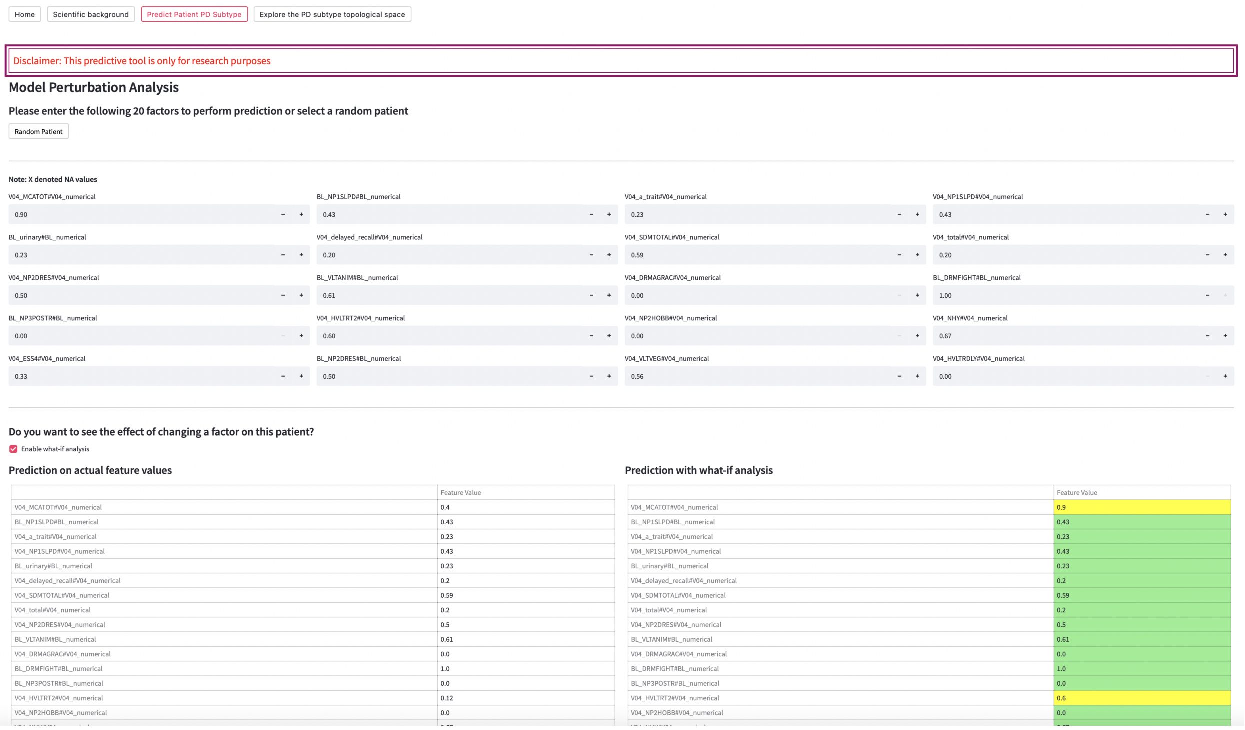


Figure 5.5: Feature importance summary plots for top features involved in PD subtype classification. The plot on the left shows the directional of each feature on model's output.


Home
Scientific background
Predict Patient PD Subtype
Explore the PD subtype topological space
Random Patient

Disclaimer: This predictive tool is only for research purposes

Model Perturbation Analysis

Please enter the following 20 factors to perform prediction or select a random patient

Note: X denoted NA values

V04_MCATOT#V04_numerical
0.90

BL_NP1SLPD#BL_numerical
0.43

V04_a_trait#V04_numerical
0.23

V04_NP1SLPD#V04_numerical
0.43

BL_urinary#BL_numerical
0.23

V04_delayed_recall#V04_numerical
0.20

V04_SDMTOTAL#V04_numerical
0.59

V04_total#V04_numerical
0.20

V04_NP2DRES#V04_numerical
0.50

BL_VLTANIM#BL_numerical
0.61

V04_DRMAGRAC#V04_numerical
0.00

BL_DRMFIGHT#BL_numerical
1.00

BL_NP3POSTR#BL_numerical
0.00

V04_HVLTRT2#V04_numerical
0.60

V04_NP2HOBB#V04_numerical
0.00

V04_NHY#V04_numerical
0.67

V04_ESS4#V04_numerical
0.33

BL_NP2DRES#BL_numerical
0.50

V04_VLTVEG#V04_numerical
0.56

V04_HVLTRDLY#V04_numerical
0.00

Do you want to see the effect of changing a factor on this patient?

☑ Enable what-if analysis

Prediction on actual feature values

	Feature Value
V04_MCATOT#V04_numerical	0.4
BL_NP1SLPD#BL_numerical	0.43
V04_a_trait#V04_numerical	0.23
V04_NP1SLPD#V04_numerical	0.43
BL_urinary#BL_numerical	0.23
V04_delayed_recall#V04_numerical	0.2
V04_SDMTOTAL#V04_numerical	0.59
V04_total#V04_numerical	0.2
V04_NP2DRES#V04_numerical	0.5
BL_VLTANIM#BL_numerical	0.61
V04_DRMAGRAC#V04_numerical	0.0
BL_DRMFIGHT#BL_numerical	1.0
BL_NP3POSTR#BL_numerical	0.0
V04_HVLTRT2#V04_numerical	0.12
V04_NP2HOBB#V04_numerical	0.0

Prediction with what-if analysis

	Feature Value
V04_MCATOT#V04_numerical	0.9
BL_NP1SLPD#BL_numerical	0.43
V04_a_trait#V04_numerical	0.23
V04_NP1SLPD#V04_numerical	0.43
BL_urinary#BL_numerical	0.23
V04_delayed_recall#V04_numerical	0.2
V04_SDMTOTAL#V04_numerical	0.59
V04_total#V04_numerical	0.2
V04_NP2DRES#V04_numerical	0.5
BL_VLTANIM#BL_numerical	0.61
V04_DRMAGRAC#V04_numerical	0.0
BL_DRMFIGHT#BL_numerical	1.0
BL_NP3POSTR#BL_numerical	0.0
V04_HVLTRT2#V04_numerical	0.6
V04_NP2HOBB#V04_numerical	0.0

Figure 5.6: Shows the interface that users can use to provide input the trained classifiers.

5.4 DISCUSSION

Our work showcases a practical approach to make machine learning techniques more accessible for the public. We have developed interactive web dashboards that enable users to explore and analyze the supervised and unsupervised models we have built as part of this dissertation. These interactive data-driven web applications provide valuable insights into

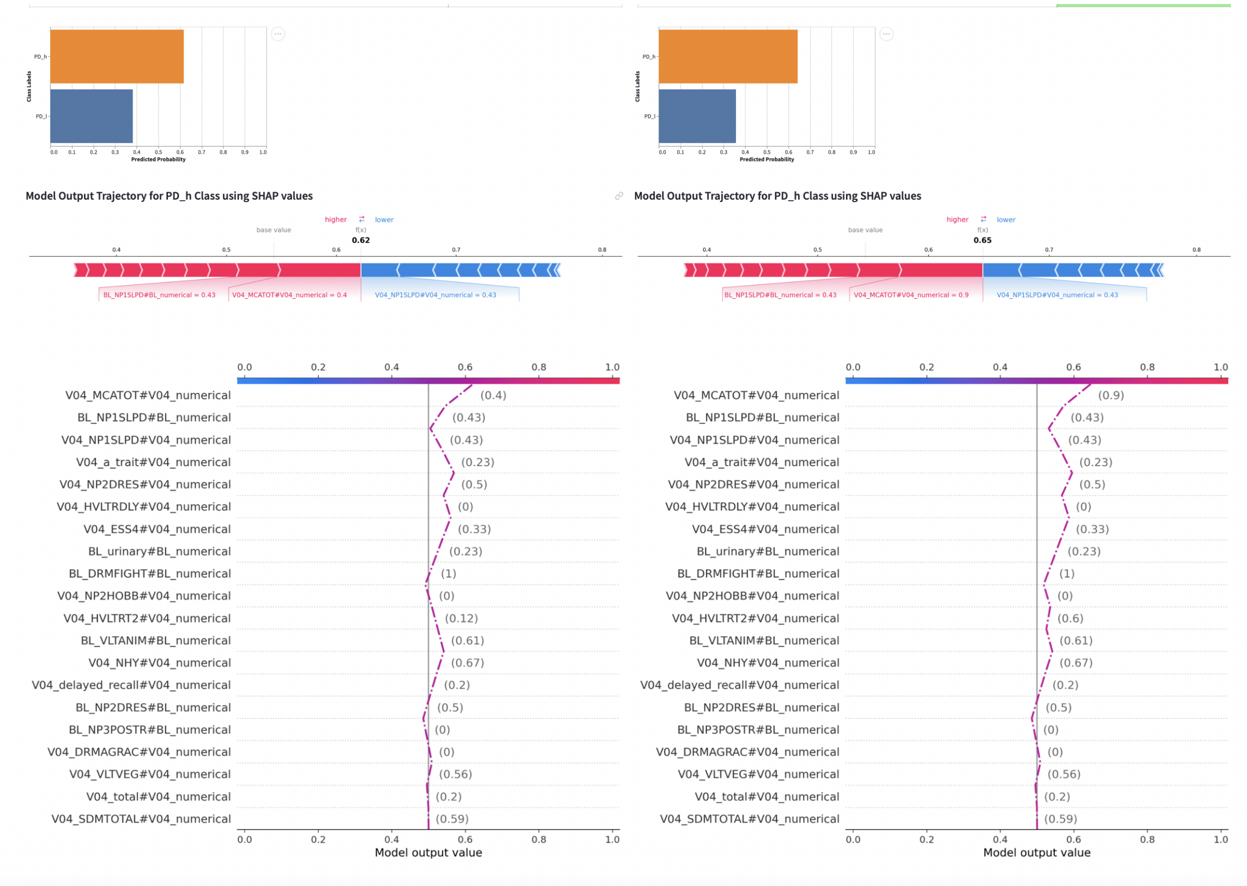
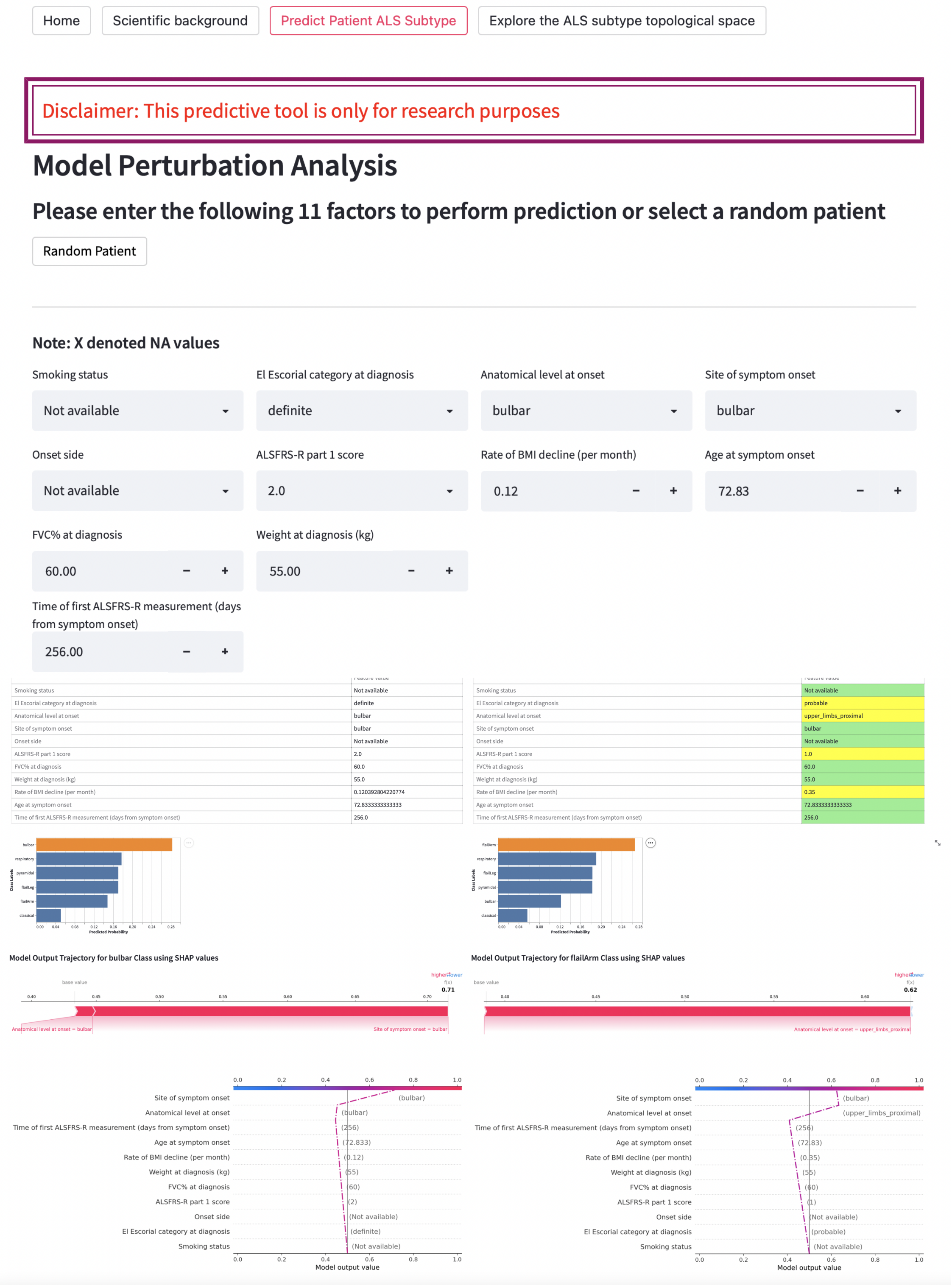


Figure 5.7: Shows the output class probabilities using trained supervised model and reasoning behind the output using SHAP force plots for PD subtype prediction.

our research results, making it easier for researchers and other healthcare professionals to understand our research outputs.

While our work represents a step forward in developing machine learning models that follow the principles of open science, it is not without its limitations. One of the major challenges we face is the high computational demand required by software tools, particularly with the ever-growing volume of large-scale healthcare datasets. Therefore, incremental learning algorithms should be developed. Analyzing health data, particularly genomics and imaging data, presents significant computational hurdles. Genomics, in particular, is widely considered to be one of the most computationally demanding big data domains, presenting challenges throughout the data life cycle, from acquisition and storage to distribution and analysis. Further, the cost of computation time in the cloud environment can be high, which highlights the importance of implementing optimized machine learning algorithms that reduce network overhead and I/O waste. Integrating more complex data modalities such as multi-omics and Magnetic Resonance Imaging (MRI) brain imaging data into our apps would be a useful addition in the future.

In the next chapter, we will provide a summary of our work and highlight the key findings of this dissertation. Additionally, we will discuss the potential directions for basic science discovery using novel biological tools based on our proof of concept results. Finally, we will conclude by presenting open research questions and future directions for further investigation.



Home
Scientific background
Predict Patient ALS Subtype
Explore the ALS subtype topological space
Random Patient

Disclaimer: This predictive tool is only for research purposes

Model Perturbation Analysis

Please enter the following 11 factors to perform prediction or select a random patient

Note: X denoted NA values

Smoking status: Not available

El Escorial category at diagnosis: definite

Anatomical level at onset: bulbar

Site of symptom onset: bulbar

Onset side: Not available

ALSFRS-R part 1 score: 2.0

Rate of BMI decline (per month): 0.12

Age at symptom onset: 72.83

FVC% at diagnosis: 60.00

Weight at diagnosis (kg): 55.00

Time of first ALSFRS-R measurement (days from symptom onset): 256.00

	Feature value
Smoking status	Not available
El Escorial category at diagnosis	definite
Anatomical level at onset	bulbar
Site of symptom onset	bulbar
Onset side	Not available
ALSFRS-R part 1 score	2.0
FVC% at diagnosis	60.0
Weight at diagnosis (kg)	55.0
Rate of BMI decline (per month)	0.120392804220774
Age at symptom onset	72.8333333333333
Time of first ALSFRS-R measurement (days from symptom onset)	256.0

	Feature value
Smoking status	Not available
El Escorial category at diagnosis	probable
Anatomical level at onset	upper_limbs_proximal
Site of symptom onset	bulbar
Onset side	Not available
ALSFRS-R part 1 score	1.0
FVC% at diagnosis	60.0
Weight at diagnosis (kg)	55.0
Rate of BMI decline (per month)	0.35
Age at symptom onset	72.8333333333333
Time of first ALSFRS-R measurement (days from symptom onset)	256.0

Model Output Trajectory for bulbar Class using SHAP values

higher ⇄ lower
f(x)
0.71
base value
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Anatomical level at onset = bulbar
Site of symptom onset = bulbar

Model Output Trajectory for flailArm Class using SHAP values

0.0
0.2
0.4
0.6
0.8
1.0
Site of symptom onset (bulbar)
Anatomical level at onset (bulbar)
Time of first ALSFRS-R measurement (days from symptom onset) (256)
Age at symptom onset (72.833)
Rate of BMI decline (per month) (0.12)
Weight at diagnosis (kg) (55)
FVC% at diagnosis (60)
ALSFRS-R part 1 score (2)
Onset side (Not available)
El Escorial category at diagnosis (definite)
Smoking status (Not available)
Model output value
0.0
0.2
0.4
0.6
0.8
1.0
Site of symptom onset (bulbar)
Anatomical level at onset (upper_limbs_proximal)
Time of first ALSFRS-R measurement (days from symptom onset) (256)
Age at symptom onset (72.83)
Rate of BMI decline (per month) (0.35)
Weight at diagnosis (kg) (55)
FVC% at diagnosis (60)
ALSFRS-R part 1 score (1)
Onset side (Not available)
El Escorial category at diagnosis (probable)
Smoking status (Not available)
Model output value

Figure 5.8: Shows the input feature interface, output class probabilities using trained supervised model and reasoning behind the output using SHAP force plots for ALS subtype prediction.

CHAPTER 6: CONCLUSIONS

6.1 INTRODUCTION

We believe our work has confirmed the significant impact that machine learning can have in advancing medical research for neurodegenerative diseases (NDDs). This work incorporated insights from prior knowledge about neurodegenerative disorders and integration of biological datasets, to build interpretable machine-learned models. We focus on four major challenges proposed by NDDs that hinders successful drug development: (i) disease heterogeneity, (ii) shared pathologies, (iii) early detection, and (iv) insufficient disease understanding. Specifically, the work leverages large datasets to address the challenges presented by the NDDs observed in medical literature to disease understanding and improvement in recruitment criteria for clinical trials.

The key innovation here is to expose the multi-modal biological data to machine learning tools – specifically neural networks – *to reduce the inefficiencies of handmade features or proxy phenotypes for neurodegenerative disorders.* We perform detailed investigation of all research analysis and deploy open-source tools for people to use and for further inspection. Our solutions demonstrate that machine learning can provide a more systematic approach to the causes and diagnosis of a disease. We believe our proposed work has the potential to influence the ML community working on healthcare problems to design specialized algorithms that work in coordination with data generated from biological research.

In Chapter 2, we deep dive into Parkinson's disease heterogeneity from the perspective of clinical signs and symptomatologies. The integration of longitudinal data was followed by vectorization and non-negative matrix factorization to develop a low-dimensional embedded space. After creating this space, we used unsupervised clustering to determine whether there were clear subtypes of disease within this space. This effort identified three distinct clinical subtypes corresponding to three groups of patients progressing at varying velocities (i.e., slow, moderate, and fast progressors). These subtypes were validated and replicated in an independent cohort. Following the successful creation of disease subtypes within a progression space, we created a baseline predictor that accurately predicted an individual patient's clinical group membership five years later. Further, we examined the predictive capability of biospecimen biomarkers at baseline and developed predictive models for early diagnosis, prognosis, and clinical trial stratification.

In Chapter 3, we explored the application of a longitudinal dimension reduction approach, Aligned-UMAP, on seven large biomedical datasets. High-dimensional longitudinal data is prevalent yet understudied in biological literature. Discovering meaningful patterns from these datasets is an important task. Though few methods are available for visualizing high dimensional longitudinal data, they are not studied extensively in real-world biological datasets. A recently developed nonlinear dimensionality reduction technique, Aligned-UMAP, analyzes sequential data. We offer insights on optimal uses for the technique and provide recommendations for best practices. Aligned-UMAP reveals time-dependent hidden patterns when color-coded with the metadata. Altogether, based on its ease of use and our evaluation of its performance on different modalities, we anticipate that Aligned-UMAP will be a valuable tool for the biomedical community. The tool is available as a web application and a GitHub repository was added to make it accessible to the users.

In Chapter 4, we discussed our work at the intersection of brain imaging and genetics data and looking multiple neurodegenerative disorders as a whole. We evaluated the potential of quantitative markers generated using machine learning techniques for Alzheimer's disease and related dementias and Parkinson's disease. We show that disease probability scores obtained from brain structural features are useful for risk stratification, prognosis prediction, and monitoring disease progression.

Chapter 5 of this dissertation presents our efforts to enhance the transparency, reproducibility, and accessibility of our research output. We discuss methods for investigating our machine learning models in great detail to ensure their validity. To facilitate more impactful research, we have deployed a cloud based web dashboard with multiple functionalities and interactive plots. Our models could serve as a potential benchmark for future research to compare against. For unsupervised models, users can explore the lower dimensional space generated from our subtype identification works and longitudinal dimensional reduction results. For supervised models, we have incorporated SHAP values and what-if analysis to open up the black box nature of trained machine learning classifiers. In summary, we believe that transparency is key to bringing machine learning into real-world healthcare settings. Further development in cloud computing is needed to make the applications faster and more cost-effective, especially with the increasing scale of high-dimensional genomics and imaging databases.

6.2 CONTRIBUTIONS

6.2.1 Publications

1. Dadu, A., Satone, V., Kaur, R. et al. Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Parkinsons Dis. 8,*

172 (2022). https://doi.org/10.1038/s41531-022-00439-z [49]

2. Dadu, Anant, et al. "Application of Aligned-UMAP to longitudinal biomedical studies." Patterns (2022). [50]

3. Prediction, prognosis and monitoring of neurodegeneration at biobank-scale via machine learning and imaging. (Dadu, Anant, et al. *In preparation* for submission at The Lancet Digital Health, 2023)

4. Faghri, Faraz, et al. Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. The Lancet Digital Health 4.5 (2022): e359-e369. [51]

5. Satone, Vipul K., et al. Predicting Alzheimer's disease progression trajectory and clinical subtypes using machine learning." bioRxiv (2019): 792432.

6. Mathew J Koretsky and others, Genetic risk factor clustering within and across neurodegenerative diseases, Brain, 2023 [52]

7. Makarious, Mary B., et al. Multi-modality machine learning predicting Parkinson's disease. npj Parkinson's Disease 8.1 (2022): 35. [53]

8. Makarious, Mary B., et al. GenoML: automated machine learning for genomics. arXiv preprint arXiv:2103.03221 (2021). [54]

9. Reilly, Luke, et al. A fully automated FAIMS-DIA proteomic pipeline for high-throughput characterization of iPSC-derived neurons. bioRxiv (2021): 2021-11. [55]

6.2.2 Web dashboards

(a) Subtypes for Amyotrophic Lateral Sclerosis (ALS): `http://bitly.ws/u5h7`

(b) Subtypes for Alzheimer's disease (AD): `http://bitly.ws/u5h3`

(c) Subtypes for Parkinson's disease (PD): `https://shorturl.at/fkvw3`

(d) Longitudinal Aligned-UMAP: `https://rb.gy/zf2xu`

An efficient and systematic approach is the key to accelerating medicinal science for NDDs therapeutic development. In this dissertation, we illustrated a framework using machine learning to address the obstacles occurring due to the complexity of neurodegenerative disorders. Our approach to PD subtype identification and prediction is a step toward precision clinical trials (1). We were able to replicate the findings in an independent cohort. A news article has discussed its potential to be incorporated into the clinical setting[5]. This robust technique has been extended to Alzheimer's disease (5). Our contribution to ALS subtype discovery offers a broad insight into clinical heterogeneity (4). This structured approach to subtype identification can improve clinical care and clinical trial design. Then, we demonstrated the applicability of a more sophisticated algorithm called Aligned-UMAP of seven different longitudinal datasets (2). It was able to capture batch effects in biological experiments (9). In addition, we provide well-documented code and a web dashboard for researchers to use on their datasets (d). As part of this thesis, we investigate genomic factors involved in Parkinson's disease using Shapley values that could provide accurate drug targets (7, 8).

Our analysis using more complex data modality, brain MRI imaging, and genomics improves the detection of PD and dementia during a patient's asymptomatic phase (3). It could be a valuable tool for early intervention when treatments may be most effective. We have used imaging and genetic data to find similarities and differences across NDDs (6). Lastly, our work on the open-science principles would bridge the gap between the use of computational research and clinical practice (a, b, c, d).

6.3 FUTURE WORK

This dissertation represents a step towards a much larger goal of developing effective drugs for neurodegenerative disorders. As part of our preliminary studies, we explored two potential research directions that align with the goals of this work. First, we discuss our efforts to understand gene functions at the cellular level, which could pave the way for novel therapeutic targets. Second, we explored the potential of extracting insights from genomic data across different neurodegenerative disorders. We conclude this section by highlighting some of the general open problems and research directions in this field that require further exploration.

[5] https://neuroderm.com/living-with-parkinson-s/info-center/machine-learning/

6.3.1 Disease associated cellular morphology

Summary: In the past two decades, mutations in several genes have been identified for NDDs, but their mechanism of action is still unknown. Understanding cellular pathology is challenging due to death and damage to neuronal cells. The development of innovative biological tools, such as iPSC and CRISPR, allows monitoring of neuronal cells in-vitro. Therefore, we propose profiling cell morphology using computer vision techniques like self-supervised learning to identify disease phenotypes at the cellular level. We will compare healthy cells with the mutated cells having risk variants identified for ALS, AD, PD, and FTD disorders. One of this work's primary research outcomes is to demonstrate ML's utility in discovering disease-specific pathology at genomic and cellular levels.

Problem And Insight: Using twin studies, heritability is estimated to be between 60% and 80% for AD and PD. This strong genetic component provides an opportunity to determine the pathophysiological processes in NDDs and identify new biological features, new prognostic/diagnostic markers, and therapeutic targets through translational genomics. Characterizing the genetic risk factors in NDDs is a primary objective; with the advent of high-throughput genomic techniques, many putative NDDs-associated loci/genes have been reported.

Cellular phenotypes have been studied for a long time, focusing on senescent cells, a type of damaged cells that are permanently withdrawn from the cell cycle [139]. Cellular senescence is widely recognized as a hallmark of aging, both as a primary causal factor in the decline of tissue homeostasis and as a consequence of other aging processes such as inflammation and DNA damage [140, 141]. Due to its critical role in disease etiology, senescence is increasingly recognized as a target for pharmaceutical intervention. Astrocytes, the largest group of glial cells in the brain, have shown senescence in Alzheimer's disease and have the potential to be novel and feasible therapeutic approaches [142, 143]. Neuronal cells might not fit into the definition of cell senescence as they are post-mitotic, but cellular phenotype can represent distinct features of stress and inflammatory response [144, 145]. However, the role of cell morphological changes in NDDs is not clearly understood. Importantly, diseased cells often exhibit an altered morphology, making them amenable to analysis with computer vision and machine learning methods.

Possible approach: Availability of cells with perturbation of a risk gene variant using CRISPR and induced pluripotent stem cell (iPSC) models presents an excellent opportunity for understanding the functions of risk gene variants and diseased phenotypes. Applying machine learning approaches to large, high-quality datasets using methods such as high-content imaging is a powerful strategy to capture novel patient- or disease-specific phenotypic

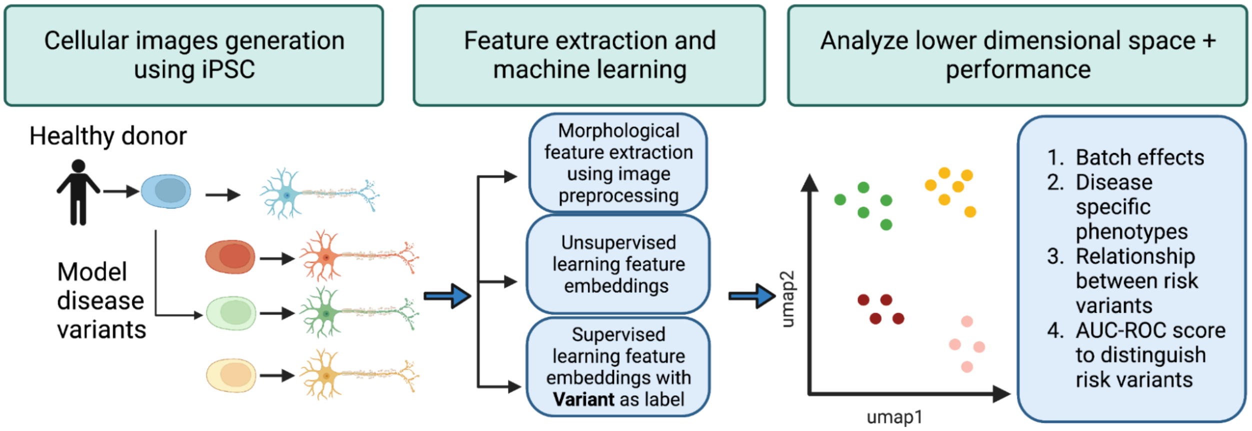

Cellular images generation using iPSC
Feature extraction and machine learning
Analyze lower dimensional space + performance
Healthy donor
Model disease variants
Morphological feature extraction using image preprocessing
Unsupervised learning feature embeddings
Supervised learning feature embeddings with Variant as label
umap2
umap1
1. Batch effects
2. Disease specific phenotypes
3. Relationship between risk variants
4. AUC-ROC score to distinguish risk variants

Figure 6.1: Workflow of analysis and model development of identification of disease cellular phenotypes.

patterns.

The workflow of the future approach could be as shown in Figure 6.1. The first step is to have the collection of the neuronal cell images captured using confocal microscopy and cell painting procedures [47]. iPSC Neurodegenerative Disease Initiative (iNDI) provides us with the cellular imaging data for FTD, ALS, AD, and PD risk variants. Then, we plan to apply three different models for feature extraction in our analysis:

- Extract morphological features such as area, shape, intensity or texture using an image processing pipeline.
- End-to-end feature learning pipeline specifically focused on unsupervised neural network techniques.
- Incorporate risk gene information into supervised neural network models. After that, we will analyze the results using classification performance metrics such as AUC-ROC and observing patterns using dimensional reduction techniques.

6.3.2 Shared mechanisms across NDDs using ML

Summary: The insight behind this work is the potential to address the overlapping clinical syndromes prevalent in NDDs by exposing biological data to machine learning techniques. Just as it is possible to understand class overlaps using ML and manifold learning techniques [146], we propose to explore using similar learning approaches on genetic risk factors identified using Genome-Wide association studies (GWAS) across NDDs. Using techniques like supervised manifold learning and ensemble learning approach, we will explore how neurodegenerative diseases overlap and provide a diagnostic tool for classifying neurodegenerative disorders using genomic data. We will use genome-wide sequencing data obtained from multiple cohorts that include Parkinson's Disease (PD) cases, Alzheimer's disease (AD) cases, Frontotemporal dementia (FTD) cases, Amyotrophic Lateral Sclerosis (ALS) cases, and Lewy body dementia (LBD) cases, with diagnostic category verified at biopsy.

Problem And Insight: All NDDs have different clinical entities concerning affected brain regions, progression, onset age, and behavior traits. However, they share similarities at different levels, such as genetic and phenotypic levels. Further, the distinction between motor and cognitive skills deficiencies across NDDs is unclear. These commonalities suggest that there might be an overlap between many of these disorders at the gene level. It can provide insight into better drug targets and lead to fundamental mechanisms behind neurodegenerative diseases. Further, understanding the differences/similarities between these disorders can provide us with better recruitment criteria for clinical trials. In addition, a gene base risk prediction model for multiple NDDs can give early disease detection and precision medicine tools. In this work, we plan to integrate multi-cohort genomic datasets for five related neurodegenerative disorders (AD, ALS, FTD, LBD, and PD), which requires adjusting for ancestry and batch effects. After that, we expose preprocessed data to machine learning techniques which can provide us with overlapping patterns of risk genes.

Possible approach: The first step is the collection of whole genome sequencing data from participants suffering from neurodegenerative disorders. Since the prevalence of NDDs such as FTD and DLB is not as high as that of AD and PD, we plan to use data from multi-cohort studies from different regions, mainly with participants of European ancestry [35]. For data harmonization, we adjust for principal components that represent population substructure to remove any population effect that can act as a confounder in our work. For the feature selection part, we use prior summary stats of GWAS research performed separately on each NDD. We plan to demonstrate results using risk variants selected at different thresholds of the p-value. After that, we trained a prediction model that can classify the NDDs using the risk variants. Once trained, we inspect the trained neural network model to study shared disease pathologies using model interpretable techniques such as Shapley Values [119] and explore low-dimensional embeddings of the penultimate layer of the neural network. We expect that observing the low dimensionality can highlight the nuances of disease overlaps and the relations between NDDs at the level of genetic risk variants. The complete workflow of the proposed model is detailed in Figure 6.2.

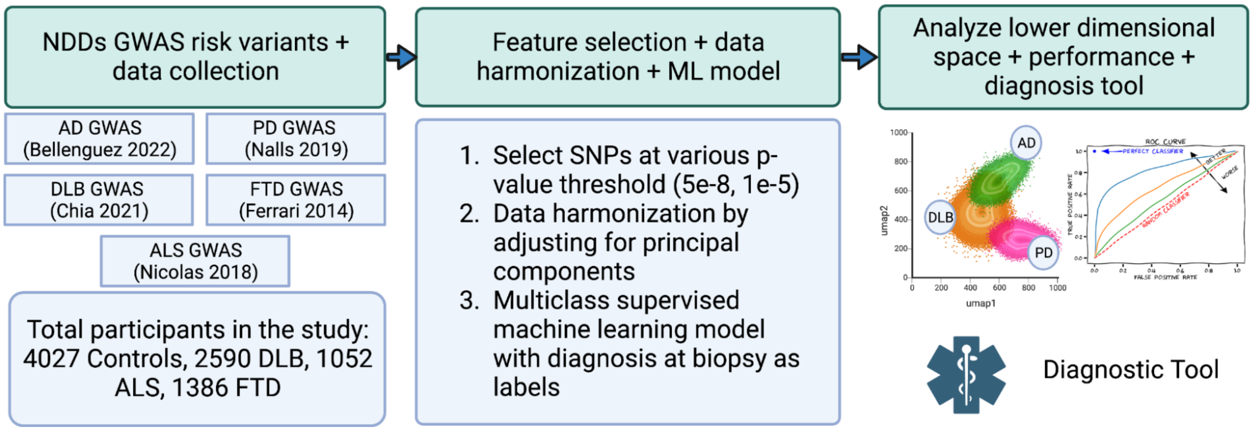

NDDs GWAS risk variants + data collection
AD GWAS (Bellenguez 2022)
PD GWAS (Nalls 2019)
DLB GWAS (Chia 2021)
FTD GWAS (Ferrari 2014)
ALS GWAS (Nicolas 2018)
Total participants in the study: 4027 Controls, 2590 DLB, 1052 ALS, 1386 FTD
Feature selection + data harmonization + ML model
1. Select SNPs at various p-value threshold (5e-8, 1e-5)
2. Data harmonization by adjusting for principal components
3. Multiclass supervised machine learning model with diagnosis at biopsy as labels
Analyze lower dimensional space + performance + diagnosis tool
AD
DLB
PD
umap2
umap1
0
200
400
600
800
1000
ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
Diagnostic Tool

Figure 6.2: Workflow of analysis and model development for extracting shared mechanisms using genomics data. The output is the hypothetical representation of the results we expect from our proposed model.

6.3.3 Open problems and research directions

There are several open problems and directions to explore further in the future:

- Data limitations
  - **Diversity:** Most of the healthcare datasets are from participants belonging to European ancestry. Diversity in the demographic makeup of healthcare data has significant implications for health equity. For example, certain diseases have a higher incidence among specific racial or ethnic groups. Failure to consider this diversity in training machine learning models can result in erroneous predictions and hazardous health outcomes [147]. Therefore, it is critical to ensure that healthcare data is representative of diverse populations to minimize healthcare disparities.
  - **Sample size**: The sample size of health data for neurodegenerative disorders (NDDs) are still not sufficient to exploit the power of advanced deep learning algorithms. More importantly there is dearth of studies having phenotypic data collected during pre-clinical stages of NDDs.
  - **Harmonization and quality control**: Machine calibrations used to collect complex healthcare datasets, such as imaging and genomics, can have a significant impact on the downstream tasks. Therefore, quality control and harmonization are crucial in data analysis pipelines.
  - **Accessiblity**: Medical data, especially from clinical settings, is often siloed and scattered across multiple institutions, making it challenging to utilize for research.

This can limit the ability of researchers and healthcare professionals to make evidence-based decisions.

  - **External validation**: In healthcare research, validation of research findings in an external cohort is a critical step to ensure the reproducibility and generalizability of the results. This is because the characteristics of the cohort used for discovery may not be representative of the larger population.

- Collaborative healthcare
  - **Biologists and data scientists:** They bring different skill sets to the table, with biologists specializing in the interpretation of biological phenomena, and data scientists applying computational tools to analyze and integrate complex data. It is particularly crucial in the analysis of complex biological data such as neuronal images or the accumulation of protein observed in PET images.
  - **Clinical trial operations and data scientists:** With regards to incorporating computational tools in setting up recruitment criteria, trial endpoints, etc., there is still a long way to go. Coordination between computer scientists and large biopharmaceutical companies conducting clinical trials is necessary to bring these tools into practice.

- Computational challenges
  - **Handling multi-modal data**: For multi-dimensional disorders such as NDDs, analyzing multi-modal is necessary. More sophisticated algorithms that can integrate data modalities such as imaging (include MRI, PET scans, DATSCAN), whole genome sequencing, and proteome data will be required.
  - **Lack of accurate labels**: Most of the datasets in neurodegenerative disorders research lacks accurate diagnostic labels. Detecting dementia accurately is a challenging task, and the current gold standard involves utilizing post-mortem brain examinations. Therefore, it is important to consider overfitting in machine learning models and more focus on unsupervised and semi-supervised learning approach is needed. It would likely be beneficial to improve MRI technology and reduce its deployment costs, alongside exploring potential improvements in other clinical approaches.
  - **Imbalanced datasets**: As more and more heterogeneity within diseases is discovered, it will become very difficult to train machine learning models for diagnos-

tic purposes due to low prevalence. Hence, methods that can handle imbalanced classes should be preferred.

  - **Storage and data size**: Development of more efficient and cost-effective storage solutions is necessary for large and complex medical datasets. This includes the use of cloud-based platforms and distributed computing systems, as well as advancements in data compression and processing techniques.

- Clinical challenges
  - **Cost-effective drugs:** The development of Alzheimer's disease drugs has made progress with Lecanemab showing some degree of clinical improvement. However, the cost of the drug is $26,500 per year, which can impose a heavy financial burden on healthcare systems. Hence, we need more research to develop cost-effective drugs for NDDs.
  - **Cost-effective biomarkers:** As research progresses and new biological data modalities are discovered, it may increase the cost of healthcare systems due to the need for additional equipment, resources, and personnel to utilize these new modalities. Therefore, we need more cost-effective biomarkers such as blood tests.

6.4 BROADER IMPACTS

This work will influence industry and academia by advocating for and enabling data-driven methodologies to enhance medical research. We provide the provide a framework to mine the ever-growing biological datasets with a comprehensive evaluation that allows smooth translation from basic science research to clinics to communities that truly benefit society. Our work has the potential to show the underpinnings of knowledge about neurodegenerative diseases that can quickly lead us to precise therapeutic targets.

Open Source Artifacts: Another primary source of value come from disseminating and maintaining well-supported open source tools. We publicized our code and data with a web application allowing researchers to inspect our methodology. Further, our analysis can work as a benchmark upon which researchers can improve with innovative algorithms and tools. For clinicians, we develop user-friendly tools that can help in their tangible decision-making concerning courses and diagnosis of neurodegenerative diseases.

REFERENCES

[1] M. R. Nelson, H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, Y. Shen, A. Floratos, P. C. Sham, M. J. Li, J. Wang et al., "The support of human genetic evidence for approved drug indications," *Nature genetics*, vol. 47, no. 8, pp. 856–860, 2015.

[2] D. Ochoa, M. Karim, M. Ghoussaini, D. G. Hulcoop, E. M. McDonagh, and I. Dunham, "Human genetics evidence supports two-thirds of the 2021 fda-approved drugs," *Nat Rev Drug Discov*, vol. 21, no. 8, p. 551, 2022.

[3] A. Dadu, A. Kumar, H. K. Shakya, S. K. Arjaria, and B. Biswas, "A study of link prediction using deep learning," in *Advanced Informatics for Computing Research: Second International Conference, ICAICR 2018, Shimla, India, July 14–15, 2018, Revised Selected Papers, Part I 2*. Springer, 2019, pp. 377–385.

[4] Z. A. Pardos, A. Dadu et al., "dafm: Fusing psychometric and connectionist modeling for q-matrix refinement," *Journal of Educational Data Mining*, vol. 10, no. 2, pp. 1–27, 2018.

[5] Z. A. Pardos and A. Dadu, "Imputing kcs with representations of problem content and context," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 2017, pp. 148–155.

[6] I. Arora, A. Dadu, M. Verma, and K. Shukla, "Random projections of fischer linear discriminant classifier for multi-class classification," in *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*. IEEE, 2016, pp. 165–169.

[7] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0268-3

[8] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Jan. 2017. [Online]. Available: https://doi.org/10.1038/nature21056

[10] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb. 2019. [Online]. Available: https://doi.org/10.1109/tpami.2018.2798607

[11] L. Gan, M. R. Cookson, L. Petrucelli, and A. R. L. Spada, “Converging pathways in neurodegeneration, from genetics to mechanisms,” *Nature Neuroscience*, vol. 21, no. 10, pp. 1300–1309, Sep. 2018. [Online]. Available: https://doi.org/10.1038/s41593-018-0237-7

[12] Y. Hou, X. Dan, M. Babbar, Y. Wei, S. G. Hasselbalch, D. L. Croteau, and V. A. Bohr, “Ageing as a risk factor for neurodegenerative disease,” *Nature Reviews Neurology*, vol. 15, no. 10, pp. 565–581, Sep. 2019. [Online]. Available: https://doi.org/10.1038/s41582-019-0244-7

[13] P. Scheltens, K. Blennow, M. M. Breteler, B. De Strooper, G. B. Frisoni, S. Salloway, and W. M. Van der Flier, “Alzheimer’s disease,” *The Lancet*, vol. 388, no. 10043, pp. 505–517, 2016.

[14] B. R. Bloem, M. S. Okun, and C. Klein, “Parkinson’s disease,” *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[15] C. H. Van Dyck, C. J. Swanson, P. Aisen, R. J. Bateman, C. Chen, M. Gee, M. Kanekiyo, D. Li, L. Reyderman, S. Cohen et al., “Lecanemab in early alzheimer’s disease,” *New England Journal of Medicine*, vol. 388, no. 1, pp. 9–21, 2023.

[16] L. Parnetti, A. Castrioto, D. Chiasserini, E. Persichetti, N. Tambasco, O. El-Agnaf, and P. Calabresi, “Cerebrospinal fluid biomarkers in parkinson disease,” *Nature Reviews Neurology*, vol. 9, no. 3, pp. 131–140, Feb. 2013. [Online]. Available: https://doi.org/10.1038/nrneurol.2013.10

[17] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. De Jong, J. Van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems et al., “Radiomics: the bridge between medical imaging and personalized medicine,” *Nature reviews Clinical oncology*, vol. 14, no. 12, pp. 749–762, 2017.

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

[19] J. Jankovic, M. McDermott, J. Carter, S. Gauthier, C. Goetz, L. Golbe, S. Huber, W. Koller, C. Olanow, I. Shoulson, M. Stern, C. Tanner, and W. W. and, “Variable expression of parkinson's disease: A base-line analysis of the DAT ATOP cohort,” *Neurology*, vol. 40, no. 10, pp. 1529–1529, Oct. 1990. [Online]. Available: https://doi.org/10.1212/wnl.40.10.1529

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

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

[22] S.-M. Fereshtehnejad, S. R. Romenets, J. B. M. Anang, V. Latreille, J.-F. Gagnon, and R. B. Postuma, “New clinical subtypes of parkinson disease and their longitudinal progression,” *JAMA Neurology*, vol. 72, no. 8, p. 863, Aug. 2015. [Online]. Available: https://doi.org/10.1001/jamaneurol.2015.0703

[23] S.-M. Fereshtehnejad, Y. Zeighami, A. Dagher, and R. B. Postuma, “Clinical criteria for subtyping parkinson’s disease: biomarkers and longitudinal progression,” *Brain*, vol. 140, no. 7, pp. 1959–1976, May 2017. [Online]. Available: https://doi.org/10.1093/brain/awx118

[24] M. Lawton, Y. Ben-Shlomo, M. T. May, F. Baig, T. R. Barber, J. C. Klein, D. M. A. Swallow, N. Malek, K. A. Grosset, N. Bajaj, R. A. Barker, N. Williams, D. J. Burn, T. Foltynie, H. R. Morris, N. W. Wood, D. G. Grosset, and M. T. M. Hu, “Developing and validating parkinson’s disease subtypes and their motor and cognitive progression,” *Journal of Neurology, Neurosurgery Psychiatry*, vol. 89, no. 12, pp. 1279–1287, July 2018. [Online]. Available: https://doi.org/10.1136/jnnp-2018-318337

[25] J. W. Vogel, A. L. Young, N. P. Oxtoby, R. Smith, R. Ossenkoppele, O. T. Strandberg, R. La Joie, L. M. Aksman, M. J. Grothe, Y. Iturria-Medina et al., “Four distinct trajectories of tau deposition identified in alzheimer’s disease,” *Nature medicine*, vol. 27, no. 5, pp. 871–881, 2021.

[26] A. L. Young, R. V. Marinescu, N. P. Oxtoby, M. Bocchetta, K. Yong, N. C. Firth, D. M. Cash, D. L. Thomas, K. M. Dick, J. Cardoso et al., “Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference,” *Nature communications*, vol. 9, no. 1, p. 4273, 2018.

[27] V. Satone, R. Kaur, A. Dadu, H. Leonard, and H. Iwaki, “Predicting alzheimer’s disease progression trajectory and clinical subtypes using machine learning.” [Online]. Available: bioRxiv.

[28] S. Byrne, P. Bede, M. Elamin, K. Kenna, C. Lynch, R. McLaughlin, and O. Hardiman, “Proposed criteria for familial amyotrophic lateral sclerosis,” *Amyotrophic Lateral Sclerosis*, vol. 12, no. 3, pp. 157–159, Jan. 2011. [Online]. Available: https://doi.org/10.3109/17482968.2010.545420

[29] J. C. Roche, R. Rojas-Garcia, K. M. Scott, W. Scotton, C. E. Ellis, R. Burman, L. Wijesekera, M. R. Turner, P. N. Leigh, C. E. Shaw et al., “A proposed staging system for amyotrophic lateral sclerosis,” *Brain*, vol. 135, no. 3, pp. 847–852, 2012.

[30] M. de Carvalho, R. Dengler, A. Eisen, J. D. England, R. Kaji, J. Kimura, K. Mills, H. Mitsumoto, H. Nodera, J. Shefner, and M. Swash, “Electrodiagnostic criteria for diagnosis of ALS,” *Clinical Neurophysiology*, vol. 119, no. 3, pp. 497–503, Mar. 2008. [Online]. Available: https://doi.org/10.1016/j.clinph.2007.09.143

[31] B. R. Brooks, "El escorial world federation of neurology criteria for the diagnosis of amyotrophic lateral sclerosis," *Journal of the Neurological Sciences*, vol. 124, pp. 96–107, July 1994. [Online]. Available: https://doi.org/10.1016/0022-510x(94)90191-0

[32] F. Faghri, F. Brunn, A. Dadu, A. Chiò, A. Calvo, C. Moglia, A. Canosa, U. Manera, R. Vasta, F. Palumbo et al., "Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study," *The Lancet Digital Health*, vol. 4, no. 5, pp. e359–e369, May 2022. [Online]. Available: https://doi.org/10.1016/s2589-7500(21)00274-0

[33] M. DeJesus-Hernandez, I. R. Mackenzie, B. F. Boeve, A. L. Boxer, M. Baker, N. J. Rutherford, A. M. Nicholson, N. A. Finch, H. Flynn, J. Adamson, N. Kouri, A. Wojtas, P. Sengdy, G.-Y. R. Hsiung, A. Karydas, W. W. Seeley, K. A. Josephs, G. Coppola, D. H. Geschwind, Z. K. Wszolek, H. Feldman, D. S. Knopman, R. C. Petersen, B. L. Miller, D. W. Dickson, K. B. Boylan, N. R. Graff-Radford, and R. Rademakers, "Expanded GGGGCC hexanucleotide repeat in noncoding region of c9orf72 causes chromosome 9p-linked FTD and ALS," *Neuron*, vol. 72, no. 2, pp. 245–256, Oct. 2011. [Online]. Available: https://doi.org/10.1016/j.neuron.2011.09.011

[34] A. E. Renton, E. Majounie, A. Waite, J. Simón-Sánchez, S. Rollinson, J. R. Gibbs, J. C. Schymick, H. Laaksovirta, J. C. Van Swieten, L. Myllykangas et al., "A hexanucleotide repeat expansion in c9orf72 is the cause of chromosome 9p21-linked als-ftd," *Neuron*, vol. 72, no. 2, pp. 257–268, 2011.

[35] R. Chia, M. S. Sabir, S. Bandres-Ciga, S. Saez-Atienzar, R. H. Reynolds, E. Gustavsson, R. L. Walton, S. Ahmed, C. Viollet, J. Ding et al., "Genome sequencing analysis identifies new loci associated with lewy body dementia and provides insights into its genetic architecture," *Nature Genetics*, vol. 53, no. 3, pp. 294–303, Feb. 2021. [Online]. Available: https://doi.org/10.1038/s41588-021-00785-3

[36] V. Barbour and R. Horton, "Mechanisms of disease," *The Lancet*, vol. 359, no. 9300, pp. 2–3, Jan. 2002. [Online]. Available: https://doi.org/10.1016/s0140-6736(02)07268-9

[37] C. Blauwendraat, M. A. Nalls, and A. B. Singleton, "The genetic architecture of parkinson's disease," *The Lancet Neurology*, vol. 19, no. 2, pp. 170–178, Feb. 2020. [Online]. Available: https://doi.org/10.1016/s1474-4422(19)30287-x

[38] C. Bellenguez, F. Küçükali, I. E. Jansen, L. Kleineidam, S. Moreno-Grau, N. Amin, A. C. Naj, R. Campos-Martin, B. Grenier-Boley, V. Andrade et al., "New insights into the genetic etiology of alzheimer's disease and related dementias," *Nature genetics*, vol. 54, no. 4, pp. 412–436, 2022.

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

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

[41] R. Ferrari, D. G. Hernandez, M. A. Nalls, J. D. Rohrer, A. Ramasamy, J. B. Kwok, C. Dobson-Stone, W. S. Brooks, P. R. Schofield, G. M. Halliday et al., "Frontotemporal dementia and its subtypes: a genome-wide association study," *The Lancet Neurology*, vol. 13, no. 7, pp. 686–699, July 2014. [Online]. Available: https://doi.org/10.1016/s1474-4422(14)70065-1

[42] G. H. Roberts, R. Partha, B. Rhead, S. C. Knight, D. S. Park, M. V. Coignet, M. Zhang, N. Berkowitz, D. A. Turrisini, M. Gaddis et al., "Expanded covid-19 phenotype definitions reveal distinct patterns of genetic association and protective effects," *Nature Genetics*, vol. 54, no. 4, pp. 374–381, 2022.

[43] B. Alipanahi, F. Hormozdiari, B. Behsaz, J. Cosentino, Z. R. McCaw, E. Schorsch, D. Sculley, E. H. Dorfman, P. J. Foster, L. H. Peng et al., "Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology," *The American Journal of Human Genetics*, vol. 108, no. 7, pp. 1217–1230, July 2021. [Online]. Available: https://doi.org/10.1016/j.ajhg.2021.05.004

[44] L. Przybyla and L. A. Gilbert, "A new era in functional genomics screens," *Nature Reviews Genetics*, vol. 23, no. 2, pp. 89–103, 2022.

[45] L. M. Shaw, M. Korecka, C. M. Clark, V. M.-Y. Lee, and J. Q. Trojanowski, "Biomarkers of neurodegeneration for diagnosis and monitoring therapeutics," *Nature reviews Drug discovery*, vol. 6, no. 4, pp. 295–303, 2007.

[46] D. M. Ramos, W. C. Skarnes, A. B. Singleton, M. R. Cookson, and M. E. Ward, "Tackling neurodegenerative diseases with genomic engineering: a new stem cell initiative from the nih," *Neuron*, vol. 109, no. 7, pp. 1080–1083, 2021.

[47] M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, and A. E. Carpenter, "Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes," *Nature Protocols*, vol. 11, no. 9, pp. 1757–1774, Aug. 2016. [Online]. Available: https://doi.org/10.1038/nprot.2016.105

[48] L. Schiff, B. Migliori, Y. Chen, D. Carter, C. Bonilla, J. Hall, M. Fan, E. Tam, S. Ahadi, B. Fischbacher et al., "Integrating deep learning and unbiased automated high-content screening to identify complex disease signatures in human fibroblasts," *Nature Communications*, vol. 13, no. 1, p. 1590, 2022.

[49] A. Dadu, V. Satone, R. Kaur, S. H. Hashemi, H. Leonard, H. Iwaki, M. B. Makarious, K. J. Billingsley, S. Bandres-Ciga, L. J. Sargent, A. J. Noyce, A. Daneshmand, C. Blauwendraat, K. Marek, S. W. Scholz, A. B. Singleton, M. A. Nalls, R. H. Campbell, and F. Faghri, "Identification and prediction of parkinson's disease subtypes and progression using machine learning in two cohorts," *npj Parkinson's Disease*, vol. 8, no. 1, Dec. 2022. [Online]. Available: https://doi.org/10.1038/s41531-022-00439-z

[50] A. Dadu, V. K. Satone, R. Kaur, M. J. Koretsky, H. Iwaki, Y. A. Qi, D. M. Ramos, B. Avants, J. Hesterman, R. Gunn, M. R. Cookson, M. E. Ward, A. B. Singleton, R. H. Campbell, M. A. Nalls, and F. Faghri, "Application of aligned-umap to longitudinal biomedical studies," *Patterns*, p. 100741, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389923000818

[51] F. Faghri, "Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study," *Lancet Digit Health*, vol. 4, p. 359– 369.

[52] M. Koretsky, C. Alvarado, M. Makarious, D. Vitale, K. Levine, A. Dadu, S. Scholz, L. Sargent, F. Faghri, and H. Iwaki, "Genetic risk factor clustering within and across neurodegenerative diseases," medRxiv. 10.1101/2022.12.01.22282945.

[53] M. B. Makarious, H. L. Leonard, D. Vitale, H. Iwaki, L. Sargent, A. Dadu, I. Violich, E. Hutchins, D. Saffo, S. Bandres-Ciga et al., "Multi-modality machine learning predicting parkinson's disease," *npj Parkinson's Disease*, vol. 8, no. 1, p. 35, 2022.

[54] M. Makarious, H. Leonard, D. Vitale, H. Iwaki, D. Saffo, L. Sargent, A. Dadu, E. Castaño, J. Carter, and M. Maleknia, "Genoml: automated machine learning for genomics," arXiv preprint arXiv:2103.03221.

[55] L. Reilly, L. Peng, E. Lara, D. Ramos, M. Fernandopulle, C. Pantazis, J. Stadler, M. Santiana, A. Dadu, and J. Iben, "A fully automated faims-dia proteomic pipeline for high-throughput characterization of ipsc-derived neurons," *bioRxiv.*

[56] A. Hughes, S. Daniel, L. Kilford, and A. Lees, "Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases," *J. Neurol. Neurosurg. Psychiatry*, vol. 55, p. 181–184.

[57] R. Postuma, "Mds clinical diagnostic criteria for parkinson's disease," *Mov. Disord*, vol. 30, p. 1591–1601.

[58] V. Satone, "Predicting alzheimer's disease progression trajectory and clinical subtypes using machine learning," preprint at. [Online]. Available: https://www.biorxiv.org/content/10.1101/792432v2

[59] M. Nalls, "Diagnosis of parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study," *Lancet Neurol*, vol. 14, p. 1002–1009.

[60] C. Sudlow, "Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med*, vol. 12, e1001779.

[61] C. Goetz, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results," *Mov. Disord*, vol. 23, p. 2129–2170.

[62] Z. Nasreddine, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *J. Am. Geriatr. Soc*, vol. 53, p. 695–699.

[63] J. Brandt, "The hopkins verbal learning test: Development of a new memory test with six equivalent forms," *Clin. Neuropsychol*, vol. 5, p. 125–142.

[64] H. Goodglass, E. Kaplan, and B. Barresi, *The assessment of aphasia and related disorders*, 3rd ed. Lippincott Williams and Wilkins.

[65] D.-I. Wechsler, *Wechsler adult intelligence scale*, 3rd ed. Psychological Corporation.

[66] A. Benton, N. Varney, and K. j. Hamsher, "A clinical test," *Arch. Neurol*, vol. 35, p. 364–367.

[67] M. Visser, J. Marinus, A. Stiggelbout, and J. Hilten, "Assessment of autonomic dysfunction in parkinson's disease: The scopa-aut," *Mov. Disord*, vol. 19, p. 1306–1312.

[68] C. Spielberger, R. Gorsuch, R. Lushene, P. Vagg, and G. Jacobs, "State-trait anxiety inventory for adults," *APA PsycTests*. [Online]. Available: https://doi.org/10.1037/t06496-000

[69] J. Yesavage and J. Sheikh, "9/geriatric depression scale (gds) recent evidence and development of a shorter version," *Clin. Gerontol*, vol. 5, p. 165–173.

[70] D. Weintraub, "Validation of the questionnaire for impulsive-compulsive disorders in parkinson's disease," *Mov. Disord*, vol. 24, p. 1461–1467.

[71] K. Stiasny-Kolster, "The rem sleep behavior disorder screening questionnaire–a new diagnostic instrument," *Mov. Disord*, vol. 22, p. 2386–2393.

[72] M. Johns, "A new method for measuring daytime sleepiness: the epworth sleepiness scale," *Sleep*, vol. 14, p. 540–545.

[73] G. McLachlan and K. Basford, "Mixture models: Inference and applications to clustering," *Marcel Dekker*, vol. 84.

[74] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat*, vol. 6, p. 461–464.

[75] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, p. 788–791.

[76] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems 13*, vol. V, p. 556–562.

[77] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, p. 1–39.

[78] L. F. M. L. Breiman, p. 5–32.

[79] G. Ke, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*, I. Guyon, Ed., p. 3146–3154.

[80] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785–794.

[81] S. Krishnagopal, R. Coelln, L. Shulman, and M. Girvan, "Identifying and predicting parkinson's disease subtypes through trajectory clustering via bipartite networks," *PLoS One*, vol. 15, e0233296.

[82] C. Marras and K. Chaudhuri, "Nonmotor features of parkinson's disease subtypes," *Mov. Disord*, vol. 31, p. 1095–1102.

[83] I. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. A Math. Phys. Eng. Sci*, vol. 374, p. 20150202.

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

[85] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. Kwok, L. Ng, F. Ginhoux, and E. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nat. Biotechnol.*

[86] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel, "A review of umap in population genetics," *J. Hum. Genet*, vol. 66, p. 85–91 10 1038 10038–020–00851–4.

[87] G. Singh, F. Mémoli, G. Carlsson, and Others, "Topological methods for the analysis of high dimensional data sets and 3d object recognition," pBG@ Eurographics 2.

[88] A. Rizvi, P. Camara, E. Kandror, T. Roberts, I. Schieren, T. Maniatis, and R. Rabadan, "Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development," *Nat. Biotechnol*, vol. 35, p. 551–560.

[89] M. Ali, M. Jones, X. Xie, and M. Williams, "Timecluster: dimension reduction applied to temporal data for visual analytics," *Vis. Comput*, vol. 35, p. 1013–1026 10 1007 00371–019–01673–.

[90] V. K. Satone, R. Kaur, A. Dadu, H. Leonard, H. Iwaki, M. Makarious, L. Sargent, A. D. N. Initiative, A. Daneshmand, S. W. Scholz et al., "Predicting alzheimer's disease progression trajectory and clinical subtypes using machine learning," *bioRxiv*, p. 792432, 2019.

[91] N. Tustison, P. Cook, A. Holbrook, H. Johnson, J. Muschelli, G. Devenyi, J. Duda, S. Das, N. Cullen, and D. Gillen, "The antsx ecosystem for quantitative biological and medical imaging," *Sci. Rep*, vol. 11, p. 9068 10 1038 41598–021–87564–6.

[92] A. Johnson, T. Pollard, L. Shen, L.-W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark, "Mimic-iii, a freely accessible critical care database," *Sci Data*, vol. 3, p. 160035.

[93] Y.-W. Lin, Y. Zhou, F. Faghri, M. Shaw, and R. Campbell, "Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory," *PLoS One*, vol. 14, p. 0218942.

[94] M. Filbin, A. Mehta, A. Schneider, K. Kays, J. Guess, M. Gentili, B. Fenyves, N. Charland, A. Gonye, and I. Gushterova, "Longitudinal proteomic analysis of severe covid-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions," *Cell Rep Med*, vol. 2, p. 100287.

[95] M. Strunz, L. Simon, M. Ansari, J. Kathiriya, I. Angelidis, C. Mayr, G. Tsidiridis, M. Lange, L. Mattner, and M. Yee, "Alveolar regeneration through a krt8+ transitional stem cell state that persists in human lung fibrosis," *Nat. Commun*, vol. 11, 3559, p. 10 1038 41467–020–17358–3.

[96] L. Reilly, L. Peng, E. Lara, D. Ramos, M. Fernandopulle, C. B. Pantazis, J. Stadler, M. Santiana, A. Dadu, J. Iben et al., "A fully automated faims-dia proteomic pipeline for high-throughput characterization of ipsc-derived neurons," *bioRxiv*, pp. 2021–11, 2021.

[97] J. Baron and R. Darling, "K-nearest neighbor approximation via the friend-of-a-friend principle," arXiv [math.CO].

[98] F. Faghri, S. Hashemi, H. Leonard, S. Scholz, R. Campbell, M. Nalls, and A. Singleton, "Predicting onset, progression, and clinical subtypes of parkinson disease using machine learning," *bioRxiv*, vol. 338913, p. 10 1101 338913.

[99] A. Sharma, E. Cao, V. Kumar, X. Zhang, H. Leong, A. Wong, N. Ramakrishnan, M. Hakimullah, H. Teo, and F. Chong, "Longitudinal single-cell rna sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy," *Nat. Commun*, vol. 9, p. 4931 10 1038 41467–018–07261–3.

[100] J. Adler-Milstein, A. Holmgren, P. Kralovec, C. Worzala, T. Searcy, and V. Patel, "Electronic health record adoption in us hospitals: the emergence of a digital "advanced use" divide," *J. Am. Med. Inform. Assoc*, vol. 24, p. 1142–1148.

[101] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "Lof: identifying density-based local outliers," *SIGMOD Rec*, vol. 29, p. 93–104.

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

[103] B. Lam, M. Masellis, M. Freedman, D. T. Stuss, and S. E. Black, "Clinical, imaging, and pathological heterogeneity of the alzheimer's disease syndrome," *Alzheimer's research & therapy*, vol. 5, no. 1, pp. 1–14, 2013.

[104] M. E. Murray, N. R. Graff-Radford, O. A. Ross, R. C. Petersen, R. Duara, and D. W. Dickson, "Neuropathologically defined subtypes of alzheimer's disease with distinct clinical characteristics: a retrospective study," *The Lancet Neurology*, vol. 10, no. 9, pp. 785–796, 2011.

[105] T. R. Fleming and D. L. DeMets, "Surrogate end points in clinical trials: are we being misled?" *Annals of internal medicine*, vol. 125, no. 7, pp. 605–613, 1996.

[106] M. Katsuno, K. Sahashi, Y. Iguchi, and A. Hashizume, "Preclinical progression of neurodegenerative diseases," *Nagoya journal of medical science*, vol. 80, no. 3, p. 289, 2018.

[107] J. Rasmussen and H. Langerman, "Alzheimer's disease–why we need early diagnosis," *Degenerative neurological and neuromuscular disease*, pp. 123–130, 2019.

[108] C. R. Jack Jr and D. M. Holtzman, "Biomarker modeling of alzheimer's disease," *Neuron*, vol. 80, no. 6, pp. 1347–1358, 2013.

[109] E. Pellegrini, L. Ballerini, M. d. C. V. Hernandez, F. M. Chappell, V. González-Castro, D. Anblagan, S. Danso, S. Muñoz-Maniega, D. Job, C. Pernet et al., "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 519–535, 2018.

[110] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, and X. Song, "Early detection of alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning," *Frontiers in neuroscience*, vol. 14, p. 259, 2020.

[111] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, and A. D. N. I. (ADNI), "Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network," *Frontiers in neuroscience*, vol. 13, p. 509, 2019.

[112] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, D. L. Collins, A. D. N. Initiative et al., "Prediction of alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning," *Neuroimage*, vol. 65, pp. 511–521, 2013.

[113] S. Spasov, L. Passamonti, A. Duggento, P. Lio, N. Toschi, A. D. N. Initiative et al., "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease," *Neuroimage*, vol. 189, pp. 276–287, 2019.

[114] D. Bzdok, G. Varoquaux, and E. W. Steyerberg, "Prediction, not association, paves the road to precision medicine," *JAMA psychiatry*, vol. 78, no. 2, pp. 127–128, 2021.

[115] S. T. Schwarz, T. Rittman, V. Gontu, P. S. Morgan, N. Bajaj, and D. P. Auer, "T1-weighted mri shows stage-dependent substantia nigra signal loss in parkinson's disease," *Movement Disorders*, vol. 26, no. 9, pp. 1633–1638, 2011.

[116] C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish et al., "Nia-aa research framework: toward a biological definition of alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[117] R. C. Petersen, H. J. Wiste, S. D. Weigand, J. A. Fields, Y. E. Geda, J. Graff-Radford, D. S. Knopman, W. K. Kremers, V. Lowe, M. M. Machulda et al., "Nia-aa alzheimer's disease framework: Clinical characterization of stages," *Annals of neurology*, vol. 89, no. 6, pp. 1145–1156, 2021.

[118] J. Ranstam and J. Cook, "Lasso regression," *Journal of British Surgery*, vol. 105, no. 10, pp. 1348–1348, 2018.

[119] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[120] A. Siderowf, L. Concha-Marambio, D.-E. Lafontant, C. M. Farris, Y. Ma, P. A. Urenia, H. Nguyen, R. N. Alcalay, L. M. Chahine, T. Foroud et al., "Assessment of heterogeneity among participants in the parkinson's progression markers initiative cohort using α-synuclein seed amplification: a cross-sectional study," *The Lancet Neurology*, vol. 22, no. 5, pp. 407–417, 2023.

[121] M. Abbas, T. Morland, M. Lichtenstein, Z. B. Milad, G. Finney, M. Boustani, and Y. El-Manzalawy, "Passive digital markers for alzheimer's disease and related dementia predict mild cognitive impairment," *Alzheimer's & Dementia*, vol. 18, p. e069373, 2022.

[122] Z. Kabelac, C. G. Tarolli, C. Snyder, B. Feldman, A. Glidden, C.-Y. Hsu, R. Hristov, E. Dorsey, and D. Katabi, "Passive monitoring at home: a pilot study in parkinson disease," *Digital biomarkers*, vol. 3, no. 1, pp. 22–30, 2019.

[123] J. Cosentino, B. Behsaz, B. Alipanahi, Z. R. McCaw, D. Hill, T.-H. Schwantes-An, D. Lai, A. Carroll, B. D. Hobbs, M. H. Cho et al., "Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models," *Nature Genetics*, pp. 1–9, 2023.

[124] M. Baker, "1, 500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452–454, May 2016. [Online]. Available: https://doi.org/10.1038/533452a

[125] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program)," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 7459–7478, 2021.

[126] B. J. Heil, M. M. Hoffman, F. Markowetz, S.-I. Lee, C. S. Greene, and S. C. Hicks, "Reproducibility standards for machine learning in the life sciences," *Nature Methods*, vol. 18, no. 10, pp. 1132–1135, Aug. 2021. [Online]. Available: https://doi.org/10.1038/s41592-021-01256-7

[127] B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, T. Shraddha, R. Kusko, S.-A. Sansone, W. Tong, R. D. Wolfinger, C. E. Mason, W. Jones, J. Dopazo, C. Furlanello, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, C. S. Greene, T. Broderick, M. M. Hoffman, J. T. Leek, K. Korthauer, W. Huber, A. Brazma, J. Pineau, R. Tibshirani, T. Hastie, J. P. A. Ioannidis, J. Quackenbush, and H. J. W. L. A. and, "Transparency and reproducibility in artificial intelligence," *Nature*, vol. 586, no. 7829, pp. E14–E16, Oct. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2766-y

[128] J. L. Robinson, E. B. Lee, S. X. Xie, L. Rennert, E. Suh, C. Bredenberg, C. Caswell, V. M. Van Deerlin, N. Yan, A. Yousef et al., "Neurodegenerative disease concomitant proteinopathies are prevalent, age-related and apoe4-associated," *Brain*, vol. 141, no. 7, pp. 2181–2193, 2018.

[129] P. H. Nguyen, A. Ramamoorthy, B. R. Sahoo, J. Zheng, P. Faller, J. E. Straub, L. Dominguez, J.-E. Shea, N. V. Dokholyan, A. De Simone et al., "Amyloid oligomers: A joint experimental/computational perspective on alzheimer's disease, parkinson's disease, type ii diabetes, and amyotrophic lateral sclerosis," *Chemical reviews*, vol. 121, no. 4, pp. 2545–2647, 2021.

[130] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725–726, Feb. 2018. [Online]. Available: https://doi.org/10.1126/science.359.6377.725

[131] M. Schlander, K. Hernandez-Villafuerte, C.-Y. Cheng, J. Mestre-Ferrandiz, and M. Baumann, "How much does it cost to research and develop a new drug? a systematic review and assessment," *PharmacoEconomics*, vol. 39, pp. 1243–1269, 2021.

[132] J. T. Shreve, S. A. Khanani, and T. C. Haddad, "Artificial intelligence in oncology: Current capabilities, future opportunities, and ethical considerations," *American Society of Clinical Oncology Educational Book*, vol. 42, pp. 842–851, 2022.

[133] F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, "Dual use of artificial-intelligence-powered drug discovery," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 189–191, 2022.

[134] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, "Dimensionality reduction by umap to visualize physical and genetic interactions," *Nature communications*, vol. 11, no. 1, p. 1537, 2020.

[135] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel, "A review of umap in population genetics," *Journal of Human Genetics*, vol. 66, no. 1, pp. 85–91, 2021.

[136] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, “Dimensionality reduction for visualizing single-cell data using umap,” *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[137] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[138] Y. Zoabi, S. Deri-Rozov, and N. Shomron, “Machine learning-based prediction of covid-19 diagnosis based on symptoms,” *npj digital medicine*, vol. 4, no. 1, p. 3, 2021.

[139] J. Campisi and F. d'Adda di Fagagna, “Cellular senescence: when bad things happen to good cells,” *Nature Reviews Molecular Cell Biology*, vol. 8, no. 9, pp. 729–740, Sep. 2007. [Online]. Available: https://doi.org/10.1038/nrm2233

[140] S. Dodig, I. Čepelak, and I. Pavić, “Hallmarks of senescence and aging,” *Biochemia medica*, vol. 29, no. 3, pp. 483–497, Oct. 2019. [Online]. Available: https://doi.org/10.11613/bm.2019.030501

[141] R. Naylor, D. Baker, and J. Van Deursen, “Senescent cells: a novel therapeutic target for aging and age-related diseases,” *Clinical Pharmacology & Therapeutics*, vol. 93, no. 1, pp. 105–116, 2013.

[142] X. Han, T. Zhang, H. Liu, Y. Mi, and X. Gou, “Astrocyte senescence and alzheimer's disease: A review,” *Frontiers in Aging Neuroscience*, vol. 12, June 2020. [Online]. Available: https://doi.org/10.3389/fnagi.2020.00148

[143] A. Bitto, C. Sell, E. Crowe, A. Lorenzini, M. Malaguti, S. Hrelia, and C. Torres, “Stress-induced senescence in human and rodent astrocytes,” *Experimental Cell Research*, vol. 316, no. 17, pp. 2961–2968, Oct. 2010. [Online]. Available: https://doi.org/10.1016/j.yexcr.2010.06.021

[144] S. K. Dehkordi, J. Walker, E. Sah, E. Bennett, F. Atrian, B. Frost, B. Woost, R. E. Bennett, T. C. Orr, Y. Zhou, P. S. Andhey, M. Colonna, P. H. Sudmant, P. Xu, M. Wang, B. Zhang, H. Zare, and M. E. Orr, “Profiling senescent cells in human brains reveals neurons with CDKN2d/p19 and tau neuropathology,” *Nature Aging*, vol. 1, no. 12, pp. 1107–1116, Dec. 2021. [Online]. Available: https://doi.org/10.1038/s43587-021-00142-3

[145] E. Sikora, A. Bielak-Zmijewska, M. Dudkowska, A. Krzystyniak, G. Mosieniak, M. Wesierska, and J. Wlodarczyk, “Cellular senescence in brain aging,” *Frontiers in Aging Neuroscience*, vol. 13, p. 646924, 2021.

[146] L. L. Zeune, Y. E. Boink, G. van Dalum, A. Nanou, S. de Wit, K. C. Andree, J. F. Swennenhuis, S. A. van Gils, L. W. Terstappen, and C. Brune, “Deep learning of circulating tumour cells,” *Nature Machine Intelligence*, vol. 2, no. 2, pp. 124–133, 2020.

[147] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, "Clinical use of current polygenic risk scores may exacerbate health disparities," *Nature genetics*, vol. 51, no. 4, pp. 584–591, 2019.