Master in High Performance Computing

Leveraging Multi-Omics and Clinical Datasets of Parkinson's Disease with Machine Learning

Supervisor(s):
Ivan Arisi,
Serafina Di Gioia, Ivan Girotto

Candidate:
Zainab Nazari

8th edition
2024

Acknowledgments

I would like to express my sincere gratitude to **Ivan Girotto** for his unwavering support, encouragement, and invaluable insights throughout my Master in High Performance Computing (MHPC). From the outset, he served as my instructor and later adviser, providing invaluable assistance and feedback that significantly contributed to the development of my thesis.

I am also deeply thankful to **Ivan Arisi** for his mentorship, encouragement, and extensive knowledge sharing during the course of my research. His expertise and guidance were instrumental in shaping the contents and methods of my thesis, and I am truly grateful for his generous support.

Additionally, I extend my heartfelt appreciation to **Serafina Di Gioia**, who transitioned from being a colleague to becoming an invaluable adviser and mentor. Her continuous support, and constructive feedback played a pivotal role in refining my thesis and enhancing its quality.

I would like to extend my heartfelt appreciation to my colleagues in MHPC program for engaging in numerous insightful discussions and sharing valuable insights throughout our study.

Last but not least, I am deeply grateful to my friends and family for their unconditional love, support and encouragement throughout my journey.

Abstract

In this thesis, I leverage the wealth of blood transcriptomic, CSF proteomics, and clinical data, including UPDRS and UPSIT scores, meticulously refining the data quality through thorough preprocessing. Employing a progressive feature selection technique, I pinpoint the most crucial genes, and proteins associated with Parkinson's disease. Subsequently, I deploy a boosting algorithm to construct a diagnostic framework centered around these identified genes and proteins. Additionally, I conduct an in-depth analysis of UPDRS and UPSIT datasets from PPMI, providing a comprehensive comparison. This holistic approach facilitates a more robust understanding of Parkinson's disease, offering insights for enhanced diagnostic and treatment strategies.

The project is conducted in collaboration with **Rita Levi-Montalcini European Brain Research Institute (EBRI)**.

Contents

Chapter 1

Introduction

The rising occurrence of Parkinson's disease and its profound health repercussions have prompted widespread research endeavors aimed at discovering efficacious treatments and early detection methodologies.

The ever-expanding volume of data associated with this pursuit, especially the assertive integration of artificial intelligence, particularly machine learning (ML), marks a noteworthy trend. A plethora of data resources is now accessible, facilitating in-depth investigations into the progression of the disease. These resources are instrumental not only in comprehending the disease's evolution but also in identifying pivotal biomarkers crucial for early diagnosis.

A recent breakthrough in the alpha-synuclein seed amplification assay (αSyn-SAA) has emerged as a promising avenue for early detection and disease monitoring [1]. Despite its potential, such diagnostic methods often necessitate invasive procedures and clinical assessments. In contrast, the well-established approach of blood transcriptomics analysis presents a non-invasive alternative, underscoring its potential to redefine the landscape of Parkinson's disease diagnosis.

Over the past decade, the Parkinson's Progression Markers Initiative (PPMI) dataset has played a pivotal role in examining the long-term progression of Parkinson's disease [2]. Utilizing ML for analyzing imaging, clinical, genetic, and multi-omics data from PPMI has proven to be crucial in advancing our understanding. In this thesis, we delve into the analysis of RNA (Ribonucleic acid) sequencing as a non-invasive approach to studying Parkinson's disease. Despite the invasive nature of obtaining cerebrospinal fluid, we capitalize on available proteomics Cerebrospinal Fluid (CSF) data from individuals with Parkinson's disease or healthy subjects at baseline, treating these datasets separately.

Our exploration has yielded notable results, employing a data glossary alongside our ML algorithm, particularly in the analysis of RNA-seq data. These findings

underscore the significance of advanced data processing and ML techniques in extracting meaningful patterns and insights from complex datasets. Additionally, we analyzed motor scores, known as UPDRS, as well as the UPSIT dataset. Therefore, our thesis is structured to discuss data extraction , preprocessing , the boosting algorithm we used and the metrics to evaluate the result and revealing the result for rna seq cfs proteomics, updrs as well as upsit. we discuss the strength of each data and how our algorithm is strutured to obtain the result and what improvement is left.

This thesis covers data extraction, preprocessing, and algorithm details. We evaluate model performance and leverage diverse datasets (RNA-seq, CSF proteomics, UPDRS, UPSIT). Examining each data source reveals its unique strengths. Finally, we analyze the results, highlighting achievements and areas for improvement.

Chapter 2 provides a panoramic overview, delving into theoretical background of subjects. Section 2.1 conducts a comprehensive review of Parkinson's disease, unraveling its origins and key features. Section 2.2 navigates the landscape of RNA-seq technology, explicating its theoretical and practical foundations.The following section, 2.3, delves into cerebrospinal fluid proteomics, providing a concise overview of its biological dimensions.

Chapter 3 focuses on data extraction and preprocessing. In section 3.1, we delve into the intricacies of the Parkinson's Progression Markers Initiative (PPMI), exploring its data repository structure and evaluating various datasets. Section 3.1.1 meticulously examines RNA-seq data from PPMI, detailing preprocessing steps, the criteria adopted for subject filtering, and relevant data empowerment methods essential for patient diagnosis. Section 3.1.2, mirroring the RNA-seq segment, encompasses data acquisition, preprocessing pipelines, and empowerment methods. The final section 3.1.3 briefly outlines the acquisition and preprocessing of UPDRS and the University of Pennsylvania Smell Identification Test (UPSIT) data.

Chapter 4 delves into the realm of ML algorithms, evaluation metrics, and the critical role of High-Performance Computing (HPC). The chapter unfolds with 4.1, which illuminates our ML algorithm, detailing the feature extraction methodology aimed at optimizing predictive performance. Transitioning to 4.1.1, we delve into the prowess of two robust ML algorithms, emphasizing their utilization of ensemble techniques. Subsequently, Section 4.2 we discuss diverse model evaluation metrics, offering insights to comprehend the obtained results. Lastly, in 4.3, we delve into the indispensability of High-Performance Computing (HPC) facilities, illustrating various scenarios where their utilization becomes imperative.

Chapter 5 is dedicated to presenting the results of our study, with each section

providing a detailed examination of the outcomes using ML evaluation metrics. In Section 5.1, we conduct an in-depth analysis of the RNA-seq data, elucidating the performance metrics and insights into the pathology of identified genes, particularly those associated with mitochondrial functions. Section 5.2 focuses on the evaluation and results derived from CSF proteomics. The final section, 5.3, presents the ML-based assessment results for both UPDRS motor scores and UPSIT data, offering a technical overview of the observed outcomes.

In the final chapter of this thesis 6, we present an overall discussion and outline potential future directions.

The provision of source code and external data in the author's GitHub repository in [3] enhances transparency and encourages collaboration in the scientific community.

Chapter 2

Background

In this chapter, firstly in 2.1 we provide a theoretical foundation for grasping Parkinson's disease, covering its definition, symptoms, and how it spreads. We then in 2.2 explain RNA-sequencing technology, shedding light on the origin of data used in our ML. Finally, in 2.3 we offer a brief introduction to cerebrospinal fluid proteomics, helping to understand better the nature of data we used in our research.

2.1 Parkinson's Disease

In 1817, James Parkinson introduced the term "shaking palsy" to characterize what we now know as Parkinsons disease (PD). PD is a neurodegenerative condition impacting the central nervous system. It is ranking as the second most prevalent neurodegenerative disorder following Alzheimer's disease that affects 2-3% of the population age $\geq$65 years old [4]. With the aging population, its prevalence is expected to rise [5], see figure 2.1. However the trend in Italy shows that there has been a large decline in the prevalence [6], which requires further investigations, some suggest that the decreasing trends in Italy might be due to the Mediterranean diet, which has been demonstrated to be related to reduced risk for PD [7], see figure 2.2. In this figure we can see the current available number of cases of Parkinson's disease per 100,000 people, in both sexes. The age followed by the age-standardized algorithm [8].

In biological point of view, PD manifests through a gradual decline in nerve cells within the **substantia nigra**, a specific region in the center of brain, see MRI image of the brain 2.3, in both healthy subject and early PD. At the time of death, this part of the brain has lost between 50% and 70% of its neurons compared to people who do not have the disease [9].This decline results in a shortage of **dopamine**, a pivotal neurotransmitter crucial for regulating movement. The primary symptoms of PD

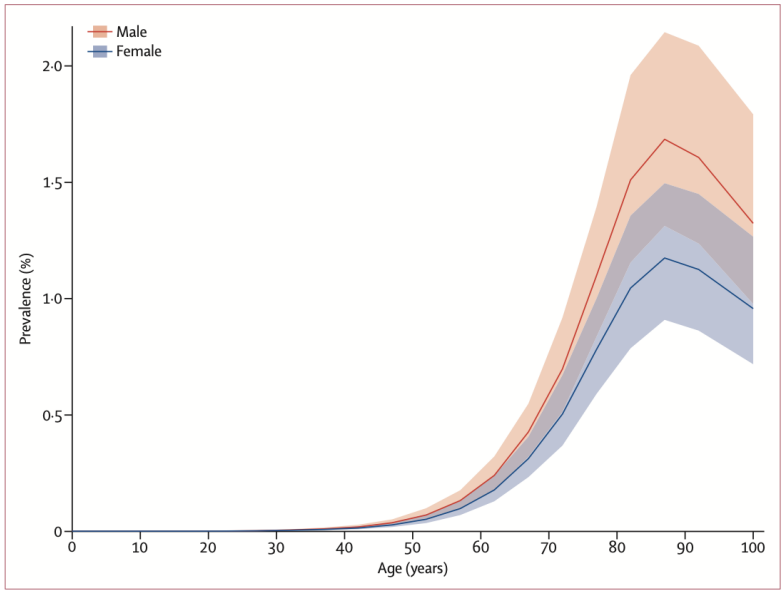

Figure 2.1: Global prevalence of PD by age and sex, 2016 [5]

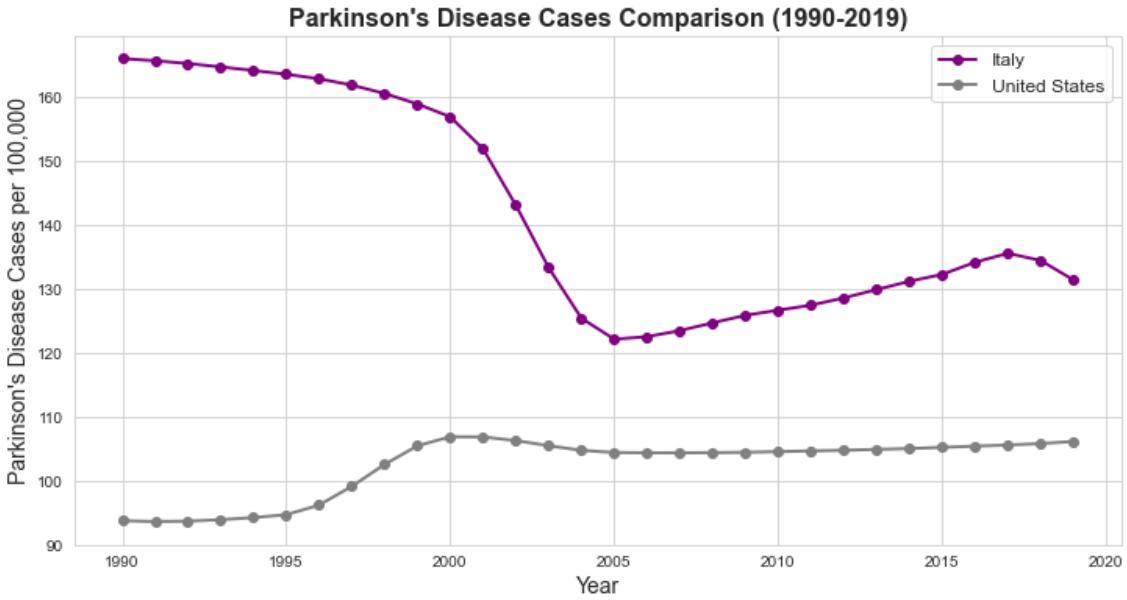

Figure 2.2: Italy versus US Parkinson's disease prevalence 1990-2019

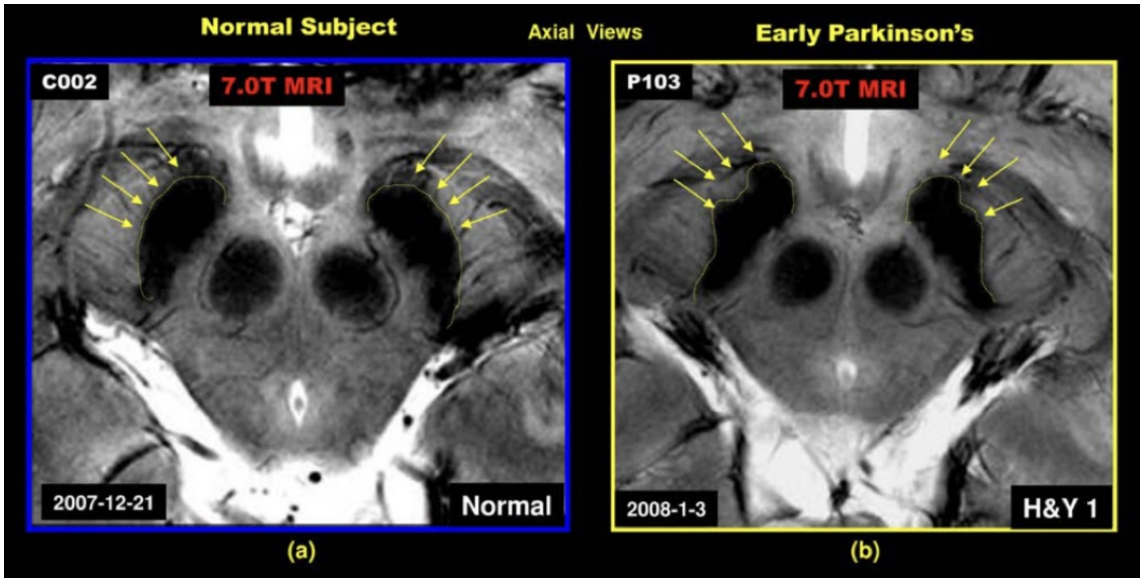

Figure 2.3: Ultra-High Field MRI : The left shows a healthy individual's substantia nigra, while the right displays a Parkinson's patient [11].

encompass several motor symptoms such as involuntary shaking, known as resting tremor which is typically observed in the hands or limbs. PD patients appear to have muscle stiffness referred to as rigidity, impeding smooth and prompt movement, and bradykinesia, characterized by slowed movement, including reduced blinking and facial expressions. Individuals with PD also experience postural instability, making balance maintenance challenging and increasing the risk of falls [10].

Normally, individuals encounter the motor symptoms of Parkinson's disease (PD) only when 50% to 70% of dopaminergic neurons have already been depleted [12]. In addition to the primary symptoms mentioned earlier, PD can exert a widespread impact on various facets of an individual's life. As the condition advances, cognitive decline becomes evident, leading to challenges in memory, thinking, and reasoning, particularly in later stages of the disease. Mental health issues, including depression, anxiety, and disruptions in sleep patterns, also emerge as significant factors influencing the overall well-being of those with PD. Speech and swallowing problems manifest as difficulties in articulation and the ingestion of food, further complicating daily life. In addition to these motor symptoms, Parkinsons disease can also cause a loss of sense of smell (anosmia), which often occurs several years before other symptoms develop. Additionally, individuals may experience pain and sensory changes, such as a sensation of burning, tingling, or numbness in the limbs. Fatigue and sleep difficulties, marked by excessive daytime sleepiness or trouble maintaining a consistent sleep pattern, further contribute to the multifaceted nature of PD. The progression of PD is typically gradual, unfolding over the course of years, with initial

symptoms becoming noticeable in middle or late life.

Parkinson's disease has an uncertain origin, likely arising from a blend of genetic and environmental elements. In terms of genetics, certain gene mutations can elevate the risk, although direct inheritance is uncommon in most cases. Environmental factors also play a role, with exposure to specific toxins, pesticides [1], and head injuries potentially contributing to a slight increase in risk. However, the evidence supporting these connections is not conclusively established.

PD lacks a cure, but treatments aim to manage symptoms and enhance quality of life. Diagnosis involves a neurological examination, imaging tests, and medication trials. Medications like levodopa address dopamine deficiency and specific symptoms. Deep brain stimulation is a surgical option for advanced cases. Therapies, including physical, occupational, and speech therapy, play a key role in managing movement, daily activities, and communication challenges associated with PD.

Continuous research efforts in PD aim to deepen our understanding of the condition, develop treatments that can modify the disease to slow or halt its progression, and enhance existing therapies. The goal is not only to grasp the intricacies of the disease but also to innovate and explore novel approaches that can offer more effective solutions for individuals affected by Parkinson's [13, 14].

2.2 RNA-Sequencing

RNA-sequencing, often abbreviated as RNA-Seq, is a technique that uses next-generation sequencing to reveal the presence and quantity of RNA molecules in a biological sample. This provides a snapshot of gene expression in the sample, also known as the transcriptome.The transcriptome refers to the complete set of RNA molecules, including messenger RNA (mRNA), non-coding RNA, and other RNA species, expressed in a cell or tissue at a specific point in time. RNA-seq provides researchers with a detailed snapshot of gene expression patterns, allowing them to quantify the abundance of different RNA molecules and identify novel transcripts.

In the past, the methods like SAGE or cDNA microarrays was used to capture these messages and then analyzed them with programming languages like R. But nowadays, a new technologies called Next-Generation Sequencing (NGS) made this process much easier and more efficient. NGS has allowed us to understand the transcriptome in a new and more complex way, and it is become the best tool for studying gene expression on a large scale. NGS is a powerful molecular biology

[1] Toxins refer to substances that can be harmful or poisonous to living organisms. These can include natural substances as well as synthetic chemicals. Pesticides are substances specifically designed to control or eliminate pests, such as insects, weeds, fungi, or rodents.

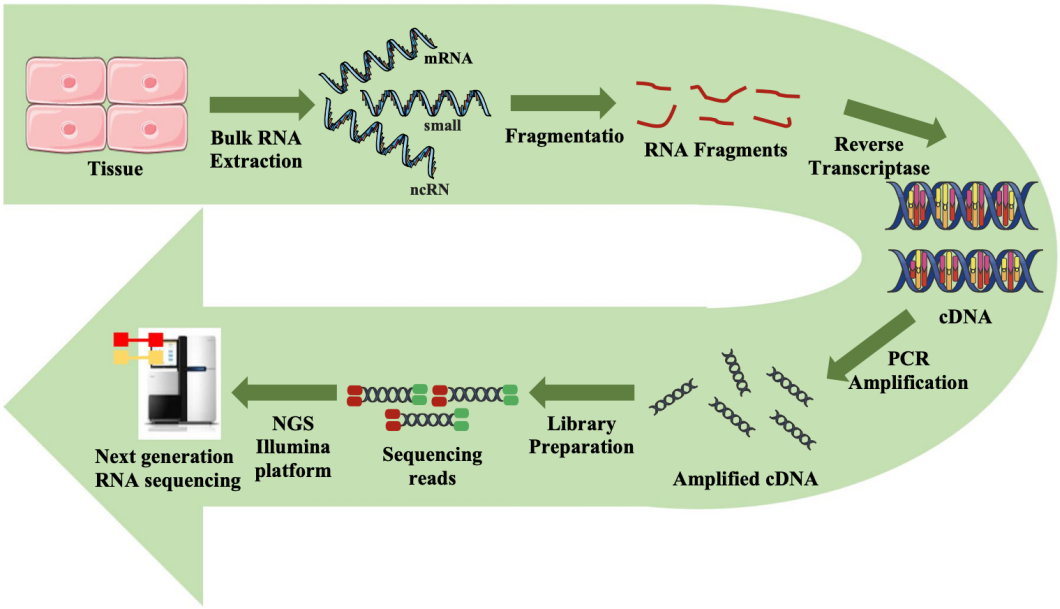

Figure 2.4: RNA-Seq technology process highlights [15]

technique that enables the comprehensive and high-throughput analysis of the entire transcriptome of an organism. This technology has revolutionized the field of genomics by offering a more precise and comprehensive understanding of the dynamic nature of gene expression, uncovering potential biomarkers, and shedding light on various biological processes.

The RNA-seq process involves several key steps. Initially, RNA is extracted from the biological sample of interest, such as cells or tissues. The extracted RNA is then fragmented to obtain smaller, more manageable pieces. Next, the fragmented RNA is converted into complementary DNA (cDNA) through reverse transcription. Subsequently, PCR amplification may occur to increase the amount of cDNA available for sequencing. After amplification, library preparation takes place, where adaptors are ligated to the ends of the cDNA fragments to facilitate sequencing. The prepared cDNA libraries are then sequenced using high-throughput sequencing platforms, such as the Illumina platform. The Illumina platform utilizes sequencing-by-synthesis technology to generate millions of short sequence reads from the cDNA fragments. These reads are subsequently aligned to the reference genome or transcriptome to quantify gene expression levels, see figure 2.4.

RNA-seq not only provides information on the abundance of known transcripts but also enables the discovery of novel transcripts, alternative splicing events, and non-coding RNAs.

RNA-seq technology offers a comprehensive snapshot of gene expressions, but

the vast amount of data it produces necessitates the use of advanced computational tools. These tools are crucial for uncovering subtle molecular patterns concealed within the broad transcriptomic landscape. The complex and extensive nature of RNA-Seq datasets poses challenges, demanding sophisticated computational methods to discern meaningful patterns and enhance our ability to predict diseases [16].

ML algorithms, renowned for their capacity to decipher intricate relationships in data, are instrumental in utilizing RNA-seq data for diagnostic purposes. They are adept at identifying nuanced gene expression patterns, pinpointing potential disease markers, and constructing models to forecast disease progression. The integration of RNA-seq data and ML not only enriches our comprehension of disease mechanisms but also paves the way for personalized medicine, customizing treatments to cater to individual needs.

In this thesis, we utilized the RNA-Seq data from PPMI. After extracting the data, and apply preprocessing steps described in 3.1.1 we employed a ML algorithm to construct a predictive model based on the available data in 4.1.

2.3 Cerebrospinal Fluid Proteomics

Cerebrospinal Fluid (CSF) is a clear fluid surrounding the brain and spinal cord, serving as a protective cushion and facilitating waste removal (refer to Figure 2.5). It acts as a valuable repository, providing profound insights into brain activities. Laden with proteins directly sourced from the brain, CSF becomes an invaluable resource for comprehending brain functions.

The samples utilized in our study were collected from participants in the Parkinson's Progression Markers Initiative (PPMI) between June 2010 and May 2019. Employing a specialized technology known as SomaScan, proteins in the CSF were quantified without access to participants' clinical details. This process employed modified aptamers called SOMAmers, ensuring reliable results [17]. The collected CSF undergoes analysis for various markers and proteins, offering potential indicators of neurological conditions, including Parkinson's disease.

In this thesis, after data extraction and preprocessing in 3.1.2 we apply ML techniques to discern specific protein patterns in CSF associated with Parkinson's disease, delving into the predictive power of CSF proteomics data in 5.2.

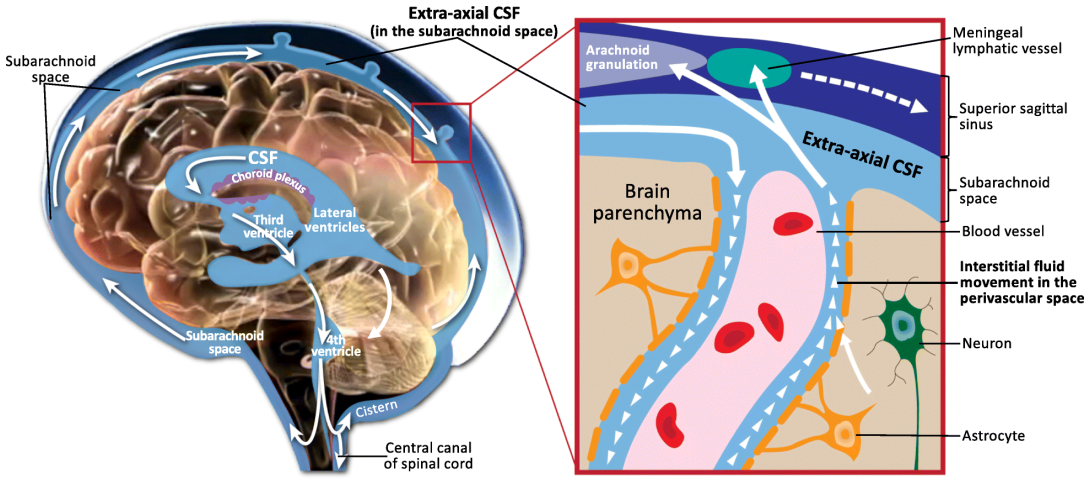

Figure 2.5: CSF, [18]

Chapter 3

Data Extraction and Preparation

In this chapter, we begin by exploring the PPMI data repository, the primary source of our data in 3.1. Subsequently, we provide a detailed overview of the data extraction and preprocessing procedures for RNA-Seq in 3.1.1, CSF in 3.1.2 proteomics, MDS-UPDRS, and UPSIT datasets in 3.1.3. These preparatory steps ensure the data is suitably formatted for integration into the ML algorithm discussed in the following chapter.

3.1 PPMI Data Repository

The Parkinson's Progression Markers Initiative (PPMI) is established in 2010 [2]. It is a significant research project sponsored by the Michael J. Fox Foundation [14]. It collects samples of DNA, RNA, plasma, serum, blood, urine, saliva, and cells from volunteers. The goal is to understand how the body changes as Parkinson's disease begins and develops. Within PPMI, data is gathered through three distinct channels: clinical, remote and online. PPMI Clinical conducts in-person clinical assessments, while PPMI Remote focuses on remote data collection, specifically targeting the pre-diagnostic phase of PD. On the other hand, PPMI Online relies on participant self-reports through a web application. Notably, PPMI Online includes a subset of information also found in the data captured by PPMI Clinical [19], see figure 3.1. In this thesis, we mainly work on PPMI clinical. In clinical channel, more than 1,500 volunteers, including those with early-stage PD, people with risk factors, and healthy volunteers, have taken part from 33 sites in 11 countries including Italy. They share data and samples for at least five years. It plans to include over 4,000 participants, including 2,000 with early signs, from almost 50 international locations, PPMI remote and online aim to arrive to 50,000 and 100,000 participant respectively[20].

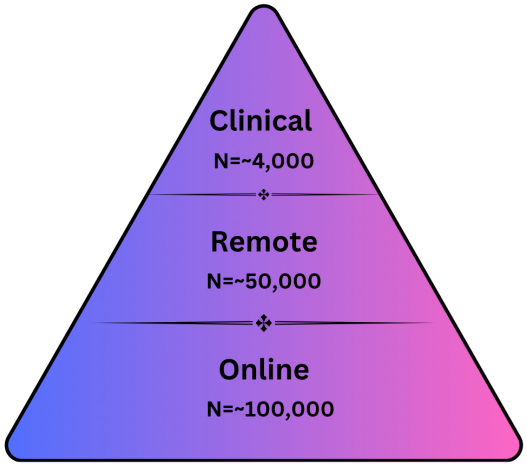

Figure 3.1: PPMI Data Channels

PPMI Clinical participants have been enrolled into one of five cohorts, namely:

- Parkinson's Disease (PD), i.e., people who have a formal diagnosis of Parkinson's disease.
- Prodromal, i.e., people who are at risk of developing PD based on clinical features, genetic variants or other biomarkers but have not been formally diagnosed. In this thesis, exclude these type of patients.
- Healthy Controls, i.e., people with no neurologic disorder and no first-degree relative with PD
- SWEDD (Scan without dopaminergic deficit). This is a small legacy cohort that we exclude.
- Early Imaging is a cohort of participants with a confirmed diagnosis of PD who were untreated and underwent additional tests including DaTscan and AV-133 imaging. In this thesis, we exclude them.

In this figure 3.2, we represent proportion of each unique cohort, and in 3.3 we can see the age distribution of participants.

PPMI is a longitudinal observational study that collects data from participants at various visits. Here's an overview of the PPMI participants by visits:

- Baseline (BL): All participants are assessed for their initial status.
- 6-Month Visit (V02): Participants return for a follow-up visit six months after the baseline.

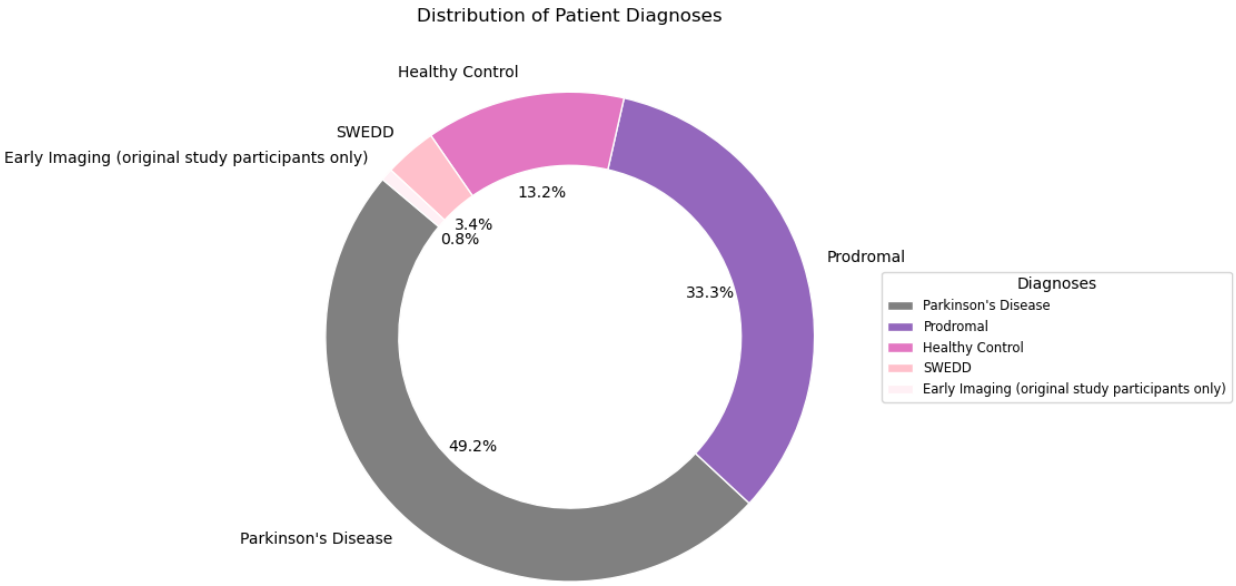

Figure 3.2: PPMI clinical participants

- Annual Visits (V04, V06, V08, etc.): Participants are assessed annually, These visits include comprehensive assessments of both motor and non-motor symptoms.
- Intermediate Visits (V05, V07, V09, etc.): These visits occur every six months between the annual visits.
- Remote Visits (R06, R08, etc.): These visits are conducted remotely and occur every six months.

The number of participants at each visit varies. Some participants may drop out of the study, while others continue to participate. The visit schedule is designed to track the progression of Parkinson's disease over time. See figure 3.4.

In this thesis, we meticulously examine baseline (BL) visits, each participants accompanied by its corresponding diagnosis either Parkinson's Disease (PD) or Healthy Control (HC) across various datasets. Subsequent sections delve into comprehensive discussions on RNA-Seq data, cerebrospinal fluid (CSF) proteomics data, as well as UPDRS and UPSIT assessments. Detailed insights are provided into the extraction process and the specific preprocessing criteria we adhere to. All the data analyzed in this study have been sourced from the Parkinson's Progression Markers Initiative (PPMI) repository.

In the subsequent sections, we delve into a detailed discussion of the data collected from individuals diagnosed with Parkinson's Disease and Healthy Controls

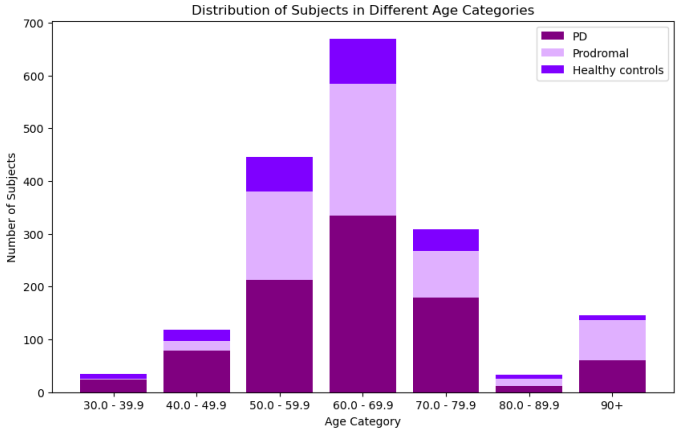

Figure 3.3: PPMI Age Distributions

during their baseline visit. This data, which includes their diagnoses, will be utilized for supervised ML assessments.

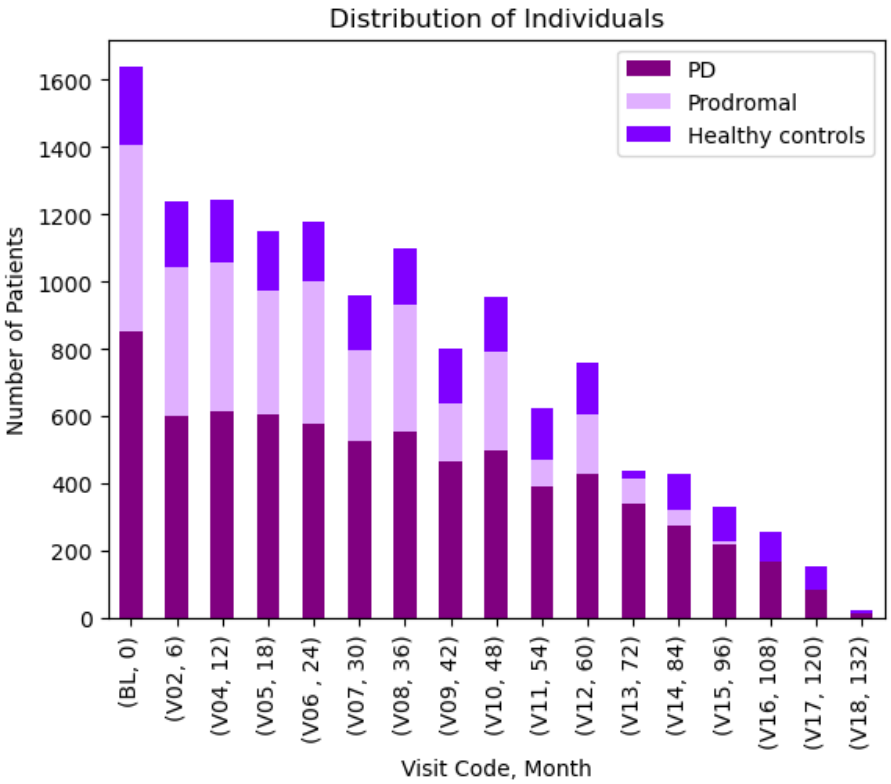

Figure 3.4: Participants by visit

3.1.1 RNA-Seq Data

In the analysis of RNA sequencing (RNA-seq) data, the quantification of gene expression hinges on two fundamental components: counts and quants.

- Counts: Represent the raw number of reads aligning to a specific gene, reflecting transcript abundance.
- Quants: Process of estimating gene expression levels based on raw sequencing data. This involves computational algorithms that transform raw read counts into meaningful expression values, taking into account factors like gene length and sequencing depth.

Our thesis utilizes counts to identify differentially expressed genes in Parkinson's disease, We provide our result in 5.1.

We retrieved the most recent RNA Sequencing Feature Counts/TPM (IR3/B38/Phases 1-2) data, version 2021-04-02, from the PPMI repository. Subsequently, we constructed a table mapping Ensembl Gene IDs to patient numbers (PATNO), encompassing 58780 genes and 1530 individuals. This was accomplished by reading,

organizing, and processing data from multiple files. To optimize the data manipulation process, we leveraged the parallel computing capabilities of Dask a library in python [21], in particular Dask DataFrame [1].

Upon the construction of the appropriate table, we implemented the following sequence of filtration and preprocessing steps:

1. Retained only individuals diagnosed as either Healthy Control or Parkinson's Disease.

2. Excluded patients with specific gene mutations (SNCA, GBA, LRRK2), particularly those with CONGBA, CONSNCA, and CONLRRK2 mutations as identified by the Consensus Committee. These patients exhibit a heightened risk of developing Parkinson's disease, more details about these gene mutations are given in Appendix B.

3. Excluded patients who were on dopaminergic medications at baseline and prior. These medications include dopamine agonists and monoamine oxidase inhibitors, both of which are commonly used in the treatment of Parkinson's disease. Dopamine agonists mimic the effects of dopamine in the brain, while monoamine oxidase inhibitors prevent the breakdown of dopamine. This information is based on a multiple treatment comparison meta-analysis studying the comparative effectiveness of dopamine agonists and monoamine oxidase type-B inhibitors for Parkinson's disease.

4. Removed duplicated gene IDs, specifically those Ensembl genes with the suffix _PAR_Y and their X transcripts.

5. Retained only genes that are either part of the 19393 protein-coding genes or the 5874 long intergenic non-coding RNAs (lincRNAs) list, as obtained from the official HGNC repository (date: 31-Jan-2024).

6. Filtered out genes with low expression levels, retaining only those genes that exhibit more than five counts in at least 10 percent of the individuals.

7. Created factors for diagnosis, sex, clinical center, and RIN using batch factor information.

8. Performed differential gene expression analysis utilizing the limma package.

[1]Dask DataFrame is a part of the Dask library and is designed to provide a parallelized and larger-than-memory alternative to pandas DataFrame

PATNO	ENSG00000000003	ENSG00000000005	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000000971	ENSG000
3000	-0.103045	-2.489666	3.992412	5.799568	3.974103	8.592125	2.144804	
3001	-1.112249	-3.836814	4.897149	5.876319	3.779341	8.469186	3.126259	
3002	1.063406	-0.122576	4.715183	5.503774	3.902711	8.225961	2.454464	
3003	-1.458928	-2.901179	4.882221	5.855794	4.065492	8.232947	3.926869	
3004	-0.597227	-1.508895	4.973190	6.080837	3.965944	8.901199	3.749207	
...	...	...	...	...	...	...	...	
4102	-0.082382	-3.631900	5.267751	6.042089	3.931937	8.010173	4.020134	
4108	-1.465083	-2.913065	4.812634	5.897239	4.226961	8.521820	3.218119	
4115	-1.408120	-3.637296	4.946118	5.971676	4.394904	8.608658	2.754131	
4136	-0.704562	-1.779543	4.174759	5.342578	3.709097	8.758651	2.533034	
4139	-0.478358	-2.873144	5.365201	6.135401	3.839386	8.517913	2.303313	

545 rows × 22033 columns

Figure 3.5: Preprocessed RNA-Seq. Data

9. Normalized factors, computed log2 counts per million, and created a design matrix with sex correction.

10. Removed batch effects using clinical center, sex, and RIN as covariates.

11. Normalized gene expression data.

After applying various preprocessing steps, we obtained a dataset consisting of 545 individuals and 22,033 genes, as depicted in Figure 3.5.

In the subsequent figure, we present a comparison of numbers of two distinct groups of individuals that were excluded from our study. This includes those with gene mutations and those using dopaminergic drugs. The intersection of these subjects is illustrated in Figure 3.6.

In the end, we apply feature scaling to gene expression data, an essential preprocessing step that normalizes or standardizes numerical our gene expression in a dataset. This scaler transforms the data by subtracting the mean and dividing by the standard deviation, ensuring that all genes share a comparable scale. This prevents particular genes from unduly influencing the learning process based on their original magnitudes. The formula for scaling a feature X is given by:

$$X_{\text{scaled}} = \frac{X - \text{mean}(X_{\text{train}})}{\text{std}(X_{\text{train}})}$$

Here, X_{scaled} denotes the scaled gene, X is the original gene, $\text{mean}(X_{\text{train}})$ signifies the mean of the training set for that gene, and $\text{std}(X_{\text{train}})$ represents the standard deviation of the training set for that gene. This process ensures that each gene has a mean of 0 and a standard deviation of 1 in the training set, maintaining uniformity across both training and test sets.

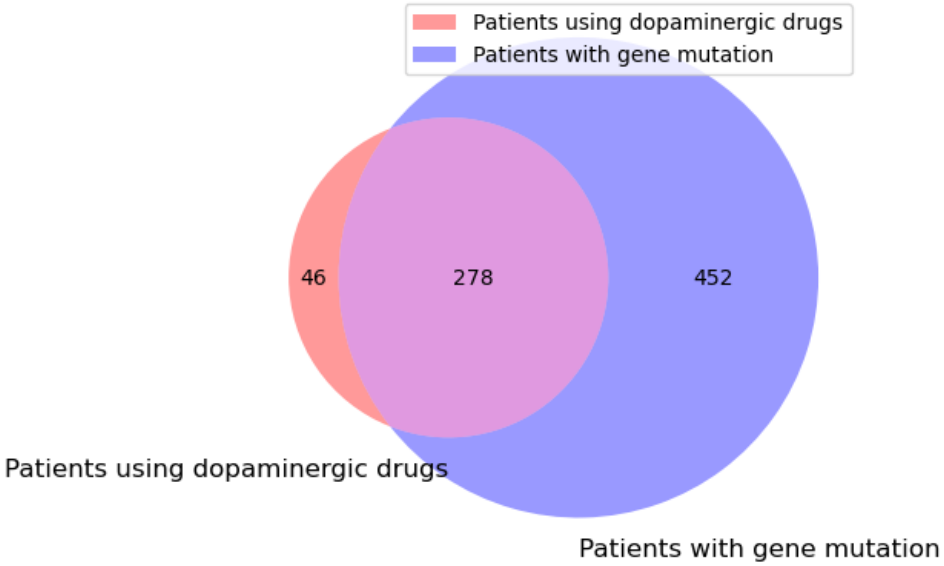

Figure 3.6: Patients excluded

3.1.2 Proteomics Data

Our analysis utilized proteomics data from the PPMI repository. We focused on extracting unique protein IDs (represented by `SOMA_SEQ_ID`) corresponding to targeted protein symbols (`TARGET_GENE_SYMBOL`). For example, $10000 - 28\_3$ corresponds to the 'CRYBB2' protein symbol, which encodes the Beta-crystallin B2 protein in humans.

These protein IDs, also referenced as `TESTNAME` in the data, have corresponding values (`TESTVALUE`) representing batch-corrected proteomic values for each patient in the CSF dataset, see Figure 3.7. These `TESTVALUE` values represent the essential proteomic data utilized for ML evaluation.
The data is initially scattered across seven files named:
Project_151_pQTL_in_CSF_{#}of_7_Batch_Corrected.csv (where # indicates a specific file number). We combined these files into a single DataFrame containing 803 subjects and 4785 proteins identified by `TESTNAME` IDs.

To ensure data quality, we excluded patients using dopaminergic drugs or having specific gene mutations (GBA, LSNCA, and LLRRK2 pathogenic variants) similar to RNA-Seq that we discussed in 3.1.1 . Following these exclusions and focusing solely on Parkinson's disease (PD) subjects and healthy controls (HC), we obtained a final dataset of 4785 proteins and 555 subjects suitable for ML analysis (see an example of the final data table in Figure 3.8).

PATNO	SEX	COHORT	CLINICAL_EVENT	TYPE	TESTNAME	TESTVALUE	UNITS	PLATEID	RUNDATE	PROJECTID	PI_NAME	PI_INSTITUTION
53595	Female	PD	BL	Cerebrospinal Fluid	5632-6_3	10.182013	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5631-83_3	6.346285	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5630-48_3	12.943830	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5629-58_3	5.891402	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5628-21_3	12.836350	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5627-53_3	7.034809	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5626-20_3	5.889332	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5624-66_3	7.657585	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5623-11_3	6.129266	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5621-64_3	5.952249	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck
53595	Female	PD	BL	Cerebrospinal Fluid	5620-13_3	10.077331	log2 RFU	P0022899	2019-08-04	151	David Stone	Merck

Figure 3.7: Proteomic gene values and IDs

PATNO	COHORT	10000-28_3	10001-7_3	10003-15_3	10006-25_3	10008-43_3	10009-2_3	10010-10_3	10011-65_3	10012-5_3	...	9984-12_3	9986-14_3
3029	0	6.881810	7.698871	6.235733	6.397376	5.898213	5.814846	6.204347	7.770759	6.309053	...	6.000616	9.258282
3963	1	6.843277	7.940987	6.253929	6.248572	5.881750	5.813261	6.203654	7.828046	6.263731	...	6.169101	8.977272
3316	0	6.833672	8.154299	6.226679	6.417975	5.950047	5.731352	6.329086	7.940774	6.328306	...	6.054259	8.856002
3124	1	6.685086	7.365837	6.227807	6.318911	6.109293	5.822314	6.023443	7.706822	6.288972	...	6.186214	8.766998
3175	1	6.984575	8.301633	6.211235	6.411799	5.994321	5.863646	6.282596	7.934654	6.257924	...	6.102391	9.922082
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3812	0	7.014804	7.469007	6.306823	6.314757	5.893445	6.001855	6.143816	7.752982	6.460554	...	6.270356	8.784713
3817	0	6.787784	7.847977	6.257174	6.412752	5.943464	5.825788	6.183518	7.829379	6.233162	...	6.128340	9.528539
3792	1	6.881702	7.962447	6.224932	6.268376	6.019198	5.847903	6.152021	7.931321	6.283648	...	6.222008	9.231704
3793	1	6.906777	7.918118	6.138533	6.500583	5.863784	5.778317	6.155346	7.687791	6.252034	...	6.098178	9.763621
3917	0	6.926580	8.053035	6.102388	6.354829	5.987164	5.856352	6.122587	7.827367	6.221964	...	6.130150	10.000907

555 rows × 4786 columns

Figure 3.8: Proteomic Gene Symbols versus Patient Numbers

3.1.3 MDS-UPDRS and UPSIT Data

The Parkinson's Progression Markers Initiative offers a wealth of clinical data for Parkinson's disease (PD) across various cohorts. However, in this section we only focus on two important ones, UPDRS and UPSIT.

The Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) plays a pivotal role in evaluating PD progression within the PPMI. This comprehensive scale comprises four distinct parts, each dedicated to assessing different facets of the disease.

- Part I focuses on Non-motor Experiences of Daily Living, encompassing items that evaluate non-motor symptoms such as sleep disturbances, fatigue, cognitive impairment, mood changes, and urinary problems. While proven useful in identifying early PD patients with cognitive decline and sleep issues, certain items may lack sensitivity for subtle changes in early PD stages.

- Part II, known as Motor Experiences of Daily Living, evaluates how PD impacts daily activities like dressing, eating, walking, and hygiene. Research on PPMI data suggests that Part II scores might not be as sensitive for early PD patients due to high item thresholds, reflecting more severe disability and potentially limiting the capture of early motor changes.

- Motor Examination (Part III) involves a clinical assessment by a trained professional, covering tremor, rigidity, bradykinesia, posture, gait, and speech. Part III scores in PPMI demonstrate good sensitivity for tracking motor progression throughout PD stages, making it a valuable tool for identifying subtle motor changes even in early PD.

- Part IV, Motor Complications, assesses complications arising from PD medication, including dyskinesia, wearing off, and freezing of gait. Part IV scores in PPMI offer insights into the development and severity of motor complications over time, aiding in treatment response monitoring and identifying patients at risk for these complications.

While the MDS-UPDRS remains a valuable tool for PD assessment in PPMI, its effectiveness varies across its parts and disease stages. Part III appears most sensitive for capturing motor changes, whereas Parts I and II may require adaptations for early PD evaluation. Therefore, in our study, we combine scores from Parts I, II, and III to evaluate PD diagnosis, see UPDRS table in 3.9. In Part I, we integrated patient questionnaires alongside clinical assessments. We focus on baseline data to assess the predictive power of these scores using ML techniques.

	PATNO	NP1RTOT	NP1PTOT	NP2PTOT	NP3TOT	Class
0	3000	3	3	0	4	HC
1	3001	0	8	2	12	PD
2	3002	3	5	15	17	PD
3	3003	1	11	6	29	PD
4	3004	0	2	0	2	HC
...	...	...	...	...	...	...
923	158359	3	0	9	10	PD
924	160040	0	2	2	25	PD
925	160231	0	2	1	23	PD
926	161236	1	1	2	13	PD
927	162929	2	12	18	28	PD

928 rows × 6 columns

Figure 3.9: MD-UPDRS Data PPMI

The University of Pennsylvania Smell Identification Test (UPSIT) data is also a crucial component in neurological studies, particularly in the context of Parkinson's disease (PD) research. UPSIT is a standardized test designed to assess an individual's ability to identify various odors, serving as an olfactory measure. As olfactory dysfunction is a common non-motor symptom of PD, UPSIT data is valuable for investigating the relationship between olfactory impairments and the progression of PD. We utilize UPSIT scores to analyze the extent of olfactory deficits in PD patients, potentially aiding in early diagnosis.

In this study, we utilize the complete dataset from the Archived UPSIT data, comprising 1905 Patient IDs and their corresponding scores for recognizing scents across four different booklets. Each score represents the recognition (1) or lack thereof (0) for ten distinct scents in each booklet: UPSITBK1, UPSITBK2, UPSITBK3, and UPSITBK4. A visual representation of the dataset is provided in Figure 3.10.

	PATNO	UPSITBK1	UPSITBK2	UPSITBK3	UPSITBK4
0	3000	9.0	10.0	6.0	10.0
2	3001	4.0	6.0	9.0	6.0
4	3002	3.0	5.0	3.0	6.0
6	3003	5.0	7.0	5.0	6.0
8	3004	10.0	10.0	8.0	8.0
...	...	...	...	...	...
2651	85236	5.0	3.0	4.0	3.0
2652	90456	5.0	5.0	4.0	5.0
2654	91097	5.0	4.0	5.0	4.0
2657	92490	3.0	1.0	3.0	3.0
2658	92834	2.0	3.0	5.0	2.0

1905 rows × 5 columns

Figure 3.10: UPSIT Data PPMI

Chapter 4

Machine Learning

In this chapter, firstly in 4.1 we explore the different ML pipelines employed across diverse datasets, each tailored to address specific challenges and optimize performance. Initially, we detail our approach for RNA-Seq and proteomics data, where our focus lies on enhancing the Area Under the Curve (AUC) score using specialized algorithms. Through iterative feature selection techniques, we refine the ML models by prioritizing the most important genes and proteins, thereby improving predictive accuracy. Subsequently, in 4.1.1 we delve into the implementation of AdaBoost and XGBoost algorithms, renowned for their prowess in boosting predictive accuracy through ensemble learning. These algorithms are meticulously applied to the various datasets, including MD-UPDRS and UPSIT, to uncover insightful patterns and relationships. Furthermore, in 4.2 we highlight the importance of comprehensive model evaluation using diverse metrics such as precision-recall curves, ROC curves, and key performance indicators like accuracy, sensitivity, and specificity. Finally, in 4.3 we emphasize the indispensable role of High-Performance Computing (HPC) in executing robust computations and optimizing ML models.

4.1 The ML Pipeline

In this section, we delve into the various ML pipelines employed across different datasets. For RNA-Seq data and proteomics data, our primary focus was on enhancing the AUC score using a specific algorithm. Additionally, we utilized another algorithm for the proteomics dataset to evaluate improvements in other ML scores. An iterative feature selection approach was implemented for both RNA-Seq gene and proteomics datasets to refine the performance of the ML models.

Initially, the correlation between features in the original datasets (comprising 22,032 genes or 4,785 proteins) and binarized patients' diagnosis (0=Healthy Con-

trol, 1=Parkinson's Disease) was evaluated by Pearson correlation index. The top 10% of features, displaying the highest absolute correlation indices, were chosen for subsequent analysis.

In the initial step, a total of 101 ML classification models distinguishing PD vs. HC were generated. This involved randomly partitioning the whole dataset 101 times for training (70%) and testing (30%) , ensuring diverse subsets for each iteration. We deployed a special boosting algorithm (AdaBoost package) with 400 decision trees, and complexity parameter of 0.0001[1]. The detals about how AdaBoost algorithm works discussed in 4.1.1. Then, the model with the highest Area Under the Curve (AUC) was selected.

Features were then prioritized based on their variable importance, and only the top 50% with the greatest importance were retained for subsequent models. In the boosting algorithm, the significance of each feature is determined by its contribution to the classification gain, which is evaluated using the Gini index within each decision tree. This importance measure is further weighted by the significance of the tree itself within the model. Specifically, in our model consisting of 400 decision trees, the feature importance is calculated individually for each tree. Subsequently, the importance of each feature is averaged across all 400 trees, by a weighted average, providing a comprehensive assessment of feature relevance within the model.

This process continued iteratively, progressively reducing the number of features while concurrently enhancing the average AUC. The iteration continued until the average AUC of the generated models reached an optimal minimum dataset size.

When the number of features dropped below 200, a refinement was introduced. Specifically, the top 75% of the most important features, rather than 50%, were selected at each step. This adjustment aimed to further optimize the procedure, retaining a higher proportion of the most influential features and discarding only the bottom 25%. When the average AUC started to decrease clearly as in Figure 5.1 following a progressive feature selection, so that the maximum top AUC was reached, the procedure was stopped. In here 4.1 we can find the flowchart of the ML algorithm for RNA-Seq dataset which we also can use with CSF proteomics data.

For the proteomics dataset (4786 proteins, 555 patients) initially the features were examined for their correlation with the binarized patients' diagnosis using the Pearson correlation index. The top 50% of features, displaying the highest absolute correlation indices, were chosen for subsequent analysis. An XGBoost model is initialized with specific hyperparameters and a seed for reproducibility. The model's

[1]The lower the complexity parameter, the more complex the trees become, resulting in an increase in the number of branches.

performance is evaluated using Stratified K-Fold cross-validation, where in each fold, the model is trained, predictions are made on the test set, and several metrics (AUC, specificity, accuracy, sensitivity) are calculated and stored. Finally, the average and standard deviation of these metrics across all folds are printed out.

Given the small size of the MD-UPDRS dataset, we refrained from employing Pearson's correlation for feature analysis and, instead, employed a default XGBoost algorithm known for its robust predictive capabilities.

Similarly, for the UPSIT data, which is also constrained by a limited number of features (only four booklets), we opted out of Pearson's correlation analysis. Instead, we constructed a ML pipeline centered around the XGBoost classifier. This entailed initializing the model with specific hyperparameters and a reproducible seed. Evaluation was carried out through Stratified K-Fold cross-validation, facilitating thorough performance assessment across multiple folds. Metrics such as ROC curve, AUC, Precision-Recall curve, and average precision were meticulously calculated and logged for each fold.

The comprehensive results for both UPDRS and UPSIT datasets are detailed in 5.3.

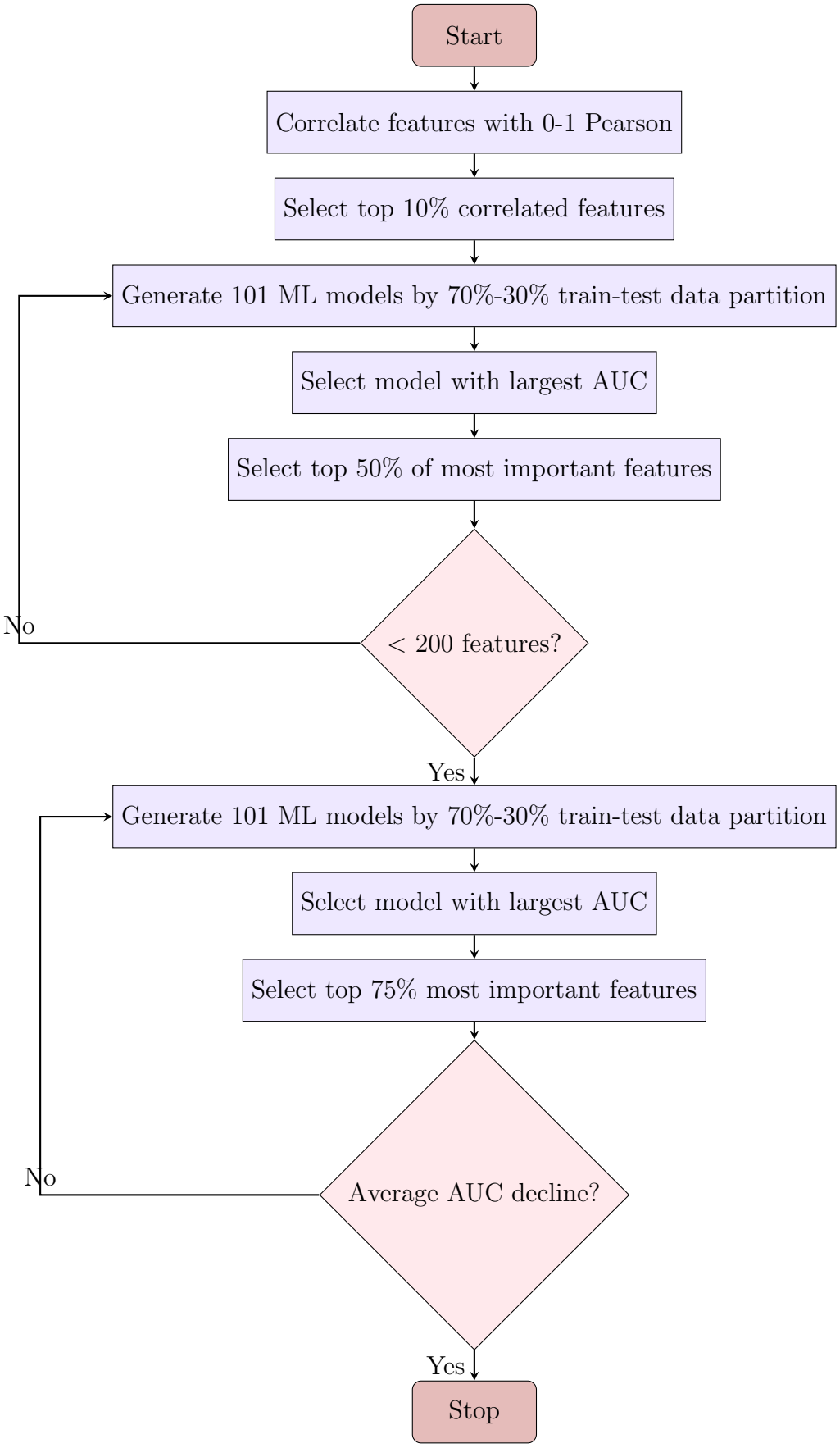

Figure 4.1: Workflow for Feature Selection and Model Refinement

4.1.1 Harnessing the Power of Ensembles

As previously mentioned, we employed both AdaBoost and XGBoost algorithms in our study. These algorithms, belonging to the ensemble learning family, are renowned for their ability to enhance predictive accuracy by leveraging the combined prediction of numerous weak learners. Let's delve into a brief description of each of these algorithms.

XGBoost, or eXtreme Gradient Boosting, has gained widespread acclaim for its efficiency and effectiveness, particularly in handling large datasets with remarkable speed and scalability. Utilizing a boosting technique, XGBoost sequentially constructs a series of weak decision trees, as exemplified in Figure 4.2, adapting subsequent trees to rectify errors made by their predecessors. Its robustness and capacity to capture intricate relationships have made XGBoost a favorite in various applications and data science competitions.

XGBoost is a powerful open-source software library providing a regularized gradient boosting framework for multiple programming languages. As an ensemble learning method, XGBoost combines the predictions of numerous weak models, typically decision trees, to create a more potent overall prediction. The iterative construction of decision trees corrects errors made by preceding ones, with the final prediction being the sum or weighted of individual tree predictions. The optimization objective function in XGBoost, defined by [22], is expressed as:

$$\text{obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \omega(f_k) \tag{4.1}$$

Here, $l(y_i, \hat{y}_i)$ represents the loss function (e.g., mean squared error for regression), $\hat{y}_i$ is the predicted value for the i-th instance, and $\omega(f_k)$ measures the complexity of the tree f_k. We utilized XGBoost to analyze the various datasets mentioned earlier. However, the compelling results obtained from analyzing the proteomics, UPDRS, and UPSIT data are presented in this thesis.

AdaBoost, or Adaptive Boosting, serves as a versatile ensemble method focused on enhancing classification accuracy. This algorithm assigns weights to misclassified data points, enabling subsequent models to prioritize more challenging instances during their iterative construction. The outcome is a powerful learner that merges the predictive strengths of numerous less powerful learners. AdaBoost's effectiveness shines in scenarios with imbalanced datasets, contributing significantly to its success in medical research and clinical studies.

For the feature selection in our RNA-Seq data, Gini index was used.

The Gini index is a metric used to evaluate the contribution of each node (fea-

ture) of the tree to the accuracy of classification. The Gini index for each decision tree node is computed as the weighted sum of Gini indices of all the child nodes, where the weights are the proportions of observations of each node. Specifically, the Gini index of each node is computed by summing the squared probabilities of each class (e.g., PD for Parkinson's disease, HC for healthy control) in the node itself. The sum of the probabilities of the two classes is always equal to 1.0 by definition. These probabilities represent the proportions of observations classified by that node.

The **Gini index** is calculated as:

$$Gini(p) = 1 - \sum_{i=1}^{2} p_i^2 \tag{4.2}$$

where p_i is the probability of an observation being classified to class i, and i can take values 1 (PD) and 2 (HC). In other words, p_1 represents the probability of an observation being classified as a Parkinson's disease case, and p_2 represents the probability of being classified as a healthy control.

In the formula, the probabilities p_i are estimated based on the proportion of samples in each class within a particular node of the tree. The goal of the decision tree algorithm is to make splits that minimize the impurity in the child nodes. The Gini index is a measure of this impurity. The lower the Gini index, the purer the node (i.e., the more it contains instances from a single class). A Gini index of 0 indicates perfect classification for the node, as only one class is present.

In this thesis, we employed AdaBoost for both the RNA-Seq and proteomics datasets for which the main algorithm is discussed in 4.1.

Both XGBoost and AdaBoost exemplify the considerable potential of ensemble learning to boost predictive accuracy and handle diverse datasets across various domains.

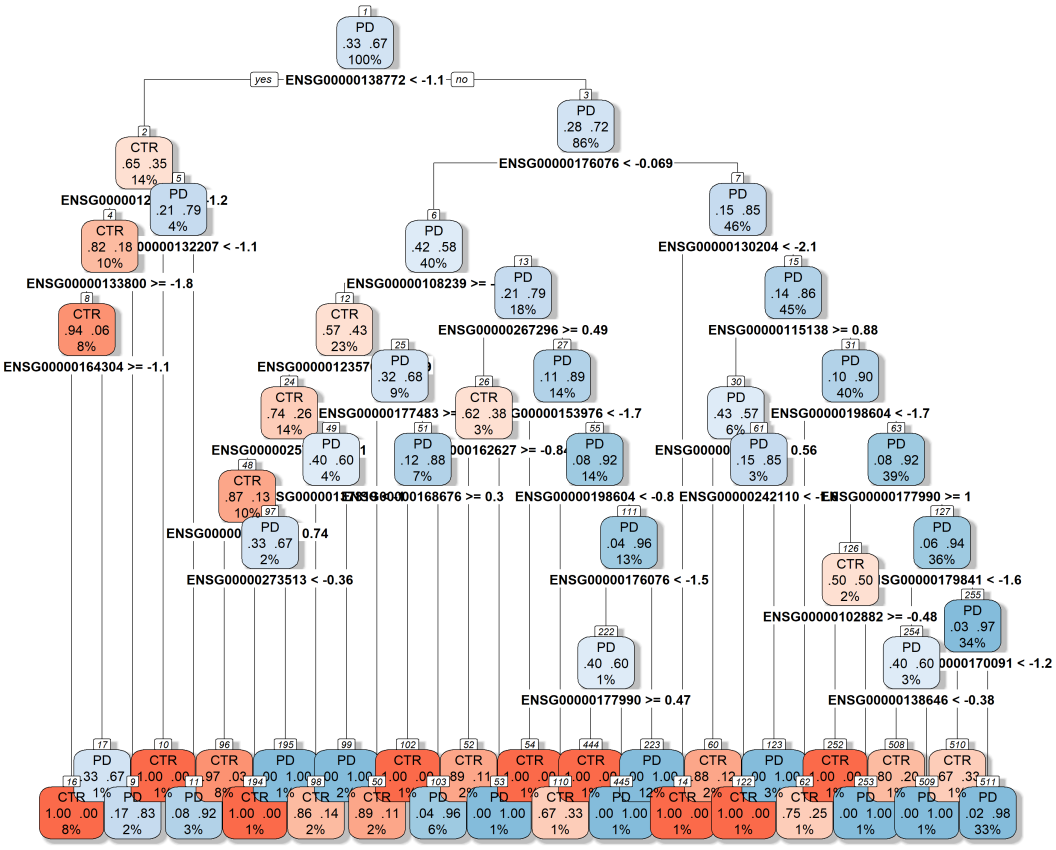

Figure 4.2: Example of Decision Tree for Diagnostic Class Prediction PD vs HC, Based on Gene Expression Prediction

4.2 Diverse Metrics for Model Assessment

In the evaluation of ML models applied to diverse datasets such as RNA-seq, proteomics CSF, UPDRS, and UPSIT data, a comprehensive understanding of various metrics to evaluate prediction performance is crucial for robust model assessment. All metrics have values in the range 0.0-1.0.

One commonly employed tool is the **precision-recall curve**, which illustrates the trade-off between precision and recall. Precision is the ratio of true positive predictions to the total predicted positives, while recall is the ratio of true positives to the total actual positives. Mathematically, precision is given by:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

And recall is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The **Receiver Operating Characteristic (ROC) curve** is another valuable tool, portraying the true positive rate against the false positive rate. Meanwhile, the confusion matrix provides a tabular representation of model performance, breaking down predictions into true positives, true negatives, false positives, and false negatives.

The **Area Under the Curve (AUC) score**, derived from the ROC curve, quantifies the model's ability to discriminate between classes.

Accuracy, a fundamental metric, measures the overall correctness of the model and is defined as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}}$$

Sensitivity, also known as recall or true positive rate, gauges the model's ability to correctly identify positive instances:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Specificity, on the other hand, assesses the model's aptitude for identifying negative instances:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Together, these metrics provide a comprehensive framework for evaluating the

performance of our ML models across diverse datasets, as outlined in Chapter 5.

4.3 The Imperative Role of HPC

The algorithms discussed in 4.1 present opportunities for improvement, largely due to the inherent complexity of the problem. RNA-Seq data, in particular, poses challenges due to its large volume of genes. Modifications to the algorithm require careful consideration, especially given the significant computational demands.

One notable challenge stems from the high dimensionality of feature spaces. As the number of features increases, so does the RAM memory and CPU time required proportionally.

In Figure 4.3, we visualize the impact of increasing the number of genes on both memory usage and CPU time for a single decision tree. In generating this plot, for each specified number of features, a simple procedure is followed. Three models are created. Firstly, the elapsed time in seconds from the start of the run is recorded, and the result is divided by three, yielding the CPU time per model. Subsequently, the memory usage is noted at the beginning of the run and after generating the three models. The difference between these memory usage values is calculated and then divided by three to obtain the RAM memory usage per model.This illustrates the scale of resources required to handle larger gene sets effectively. Memory usage increases more rapidly than CPU time and more than linearly in log-log scale. In light of the current state-of-the-art high-end hardware for High-Performance Computing , this observation emphasizes the importance of optimizing algorithms to efficiently utilize hardware resources and address the computational demands of processing RNA-Seq data. Further refinement and optimization of the algorithm is necessary to enhance its scalability and performance on modern HPC systems.

An essential application of high-performance computing lies in executing robust computations, notably with cross-validation techniques. While we initially used a random partition of genes in the RNA-Seq evaluation to enhance the AUC score, there is a compelling need to reevaluate the model using cross-validation, especially repeated stratified cross-validation. Given the extensive number of genes that need to be assessed, coupled with the iterative process of cross-validation that involves multiple folds and iterations, the computational time and memory requirements are significantly amplified. High-Performance Computing (HPC) facilities, equipped with parallel processing capabilities, serve as a crucial solution to these challenges.

In the process of evaluating and optimizing our ML models we conduct numerous simulations with varying parameters. Leveraging HPC enables swift exploration of

these options, facilitating faster experimentation and refinement of hyperparameters to enhance model efficiency.

In boosting models the main parameters to be set are the number of trees (classifiers) and the complexity cp, the latter connected to the tree depth. One of the next steps will be to use also neural network modelling, for which the number of parameters will significantly increase. In the case of PPMI datasets, a further improvement of transcritomic-based diagnostic class-prediction model will be to integrate gene expression data with single nucleotide polymorphisms (SNP) gene variants data, also provided by the PPMI database. For SNP data the number of features is of the order of millions and a pruning filtering step will be necessary, discarding highly correlated features.

Feature importance analysis is a common step of ML pipeline that usually places a considerable burden on computational resources. This is particularly true in situations with an extensive number of features. Algorithms tasked with determining variable importance contribute to increased computational complexity, demanding powerful computing facilities for accurate and timely assessments.

Conventional computing resources often face limitations in handling large-scale ML tasks. Extended processing times and potential memory overflow become significant concerns. In such cases, HPC solutions emerge as a game-changer. Parallel and serial optimization, together with GPU computing, the high-speed networks and increased memory capacity offered by HPC facilities, allow us to solve efficiently time and memory critical problems.

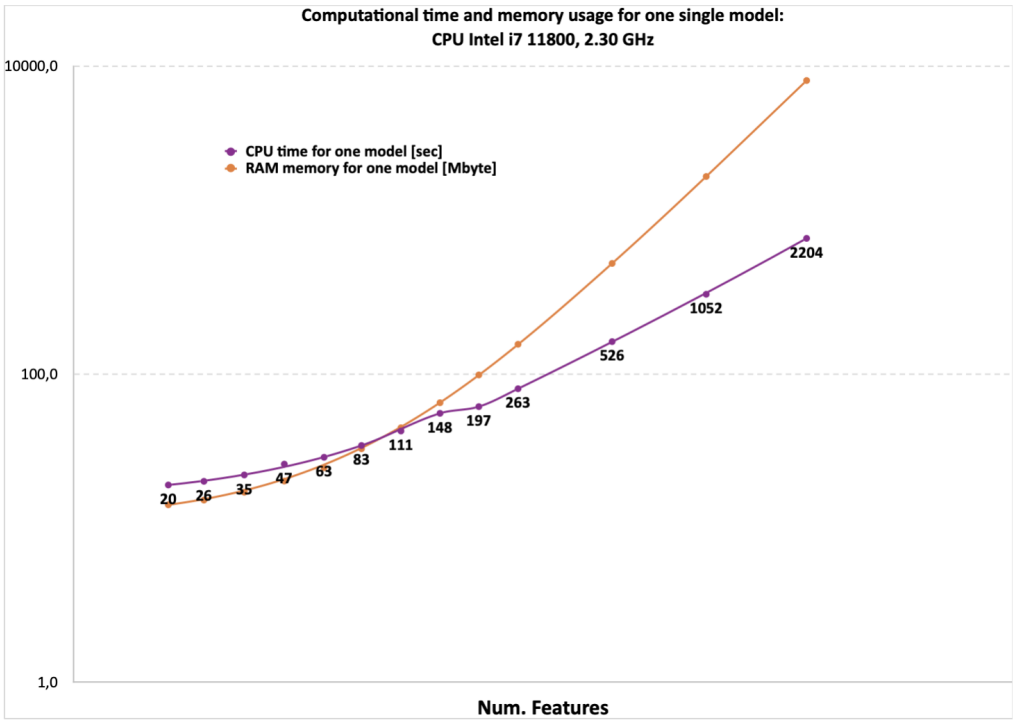

Figure 4.3: CPU time and memory usage as a function of features, in log-log scale.

Chapter 5

Results

In this chapter, we delve into the outcomes generated by our ML algorithms as outlined in Section 4.1. Initially, we examine the results derived from RNA-Seq analysis, focusing on the mean evaluation metrics across various gene sets and highlighting the findings with the highest mean AUC. Subsequently, in Section 5.2, we elaborate on the outcomes obtained through two distinct algorithms to analyze the proteomic dataset: one employing Adaboost with progressive protein selection, and the other utilizing XGBoost while retaining a larger subset of proteins. We thoroughly analyze the results using diverse metrics. Finally, in Section 5.3, we showcase the results of our ML endeavors and discuss the predictive prowess they demonstrate.

5.1 RNA-Seq Result

Our findings, with the ML pipeline detailed in 4.1, emphasize the potential of progressive machine learning (ML) technique with AdaBoost, for the diagnosis of Parkinson's disease. This is particularly significant when dealing with extensive datasets such as RNA-Seq data.

Employing a progressive ML algorithm that selectively discards genes at each step, see 4.1, we observed intriguing results, as shown for diverse metrics across different number of gene predictors in Figure 5.1. Specifically, with 148 genes, we achieved a maximum mean Area Under the ROC Curve (AUC) of 0.852 with a variability of SD = 0.029, and a single model achieving AUC of 0.926. This suggests a robust capability of this algorithm to discriminate between Parkinson's patients and healthy controls.

In Figure 5.2, we can observe a series of ROC curves generated with these genes, executed over 101 iterations, where the red curve represents the mean performance.

Furthermore, the highest mean accuracy of 0.785, observed with 111 genes, un-

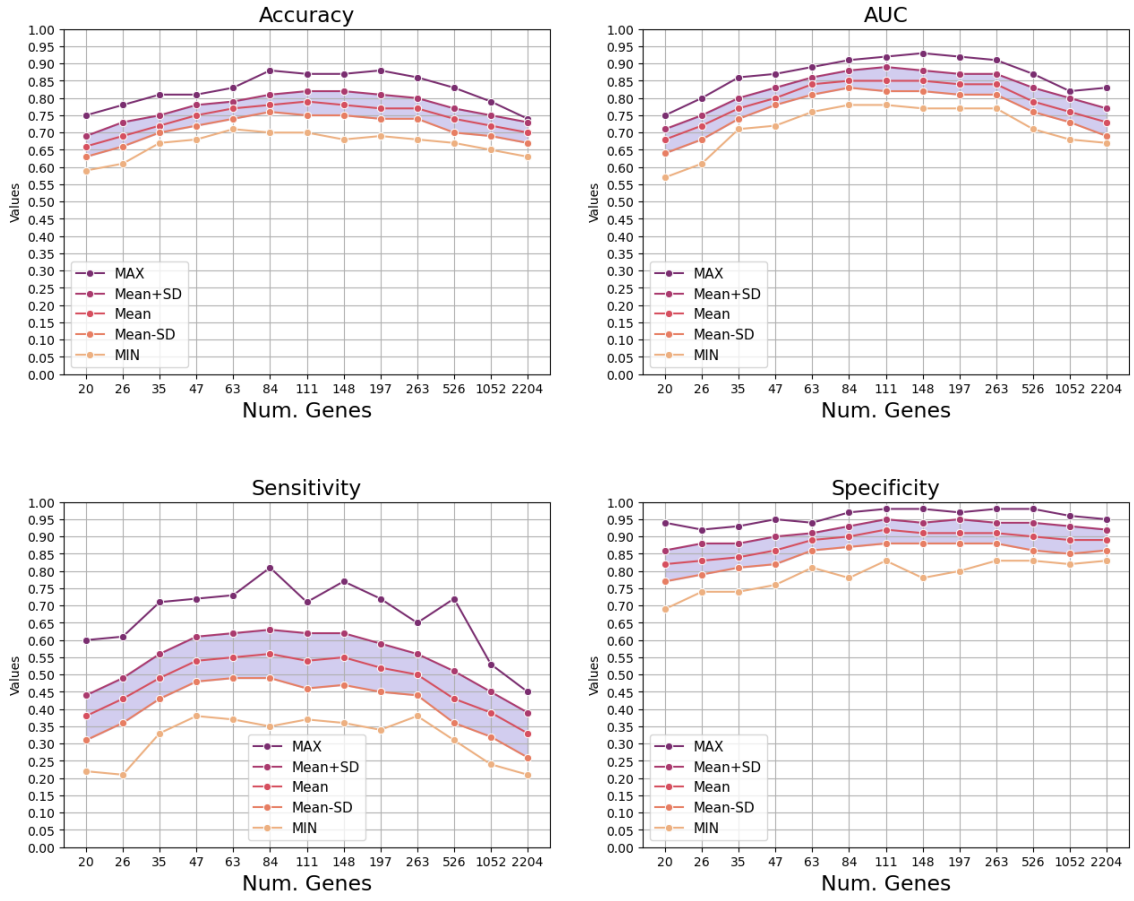

Figure 5.1: Performance Metrics of ML Models on the RNA-Seq Dataset

derscores a promising overall classification performance. However, a notable disparity arises between the maximum mean specificity (0.915) and the maximum mean sensitivity (0.563).

Our analysis highlights the significant contribution of genes associated with mitochondria on enhancing the accuracy of our machine learning model. However, subsequent findings suggest that while these genes contribute significantly to the model's performance, they may not independently regulate the development of Parkinson's Disease.

The Mitochondrion, a vital organelle found in most eukaryotic cells, plays a crucial role in cellular respiration and energy production. It is enclosed by membranes within the cell's cytoplasm and is essential for generating adenosine triphosphate (ATP), a key energy currency in cells. Research suggests a connection between mitochondrial dysfunction and Parkinson's disease. Mitochondria serve as energy sources for dopaminergic neurons in the brain, critical for motor function. Dysfunctional mitochondria may contribute to the loss of dopaminergic neurons in the

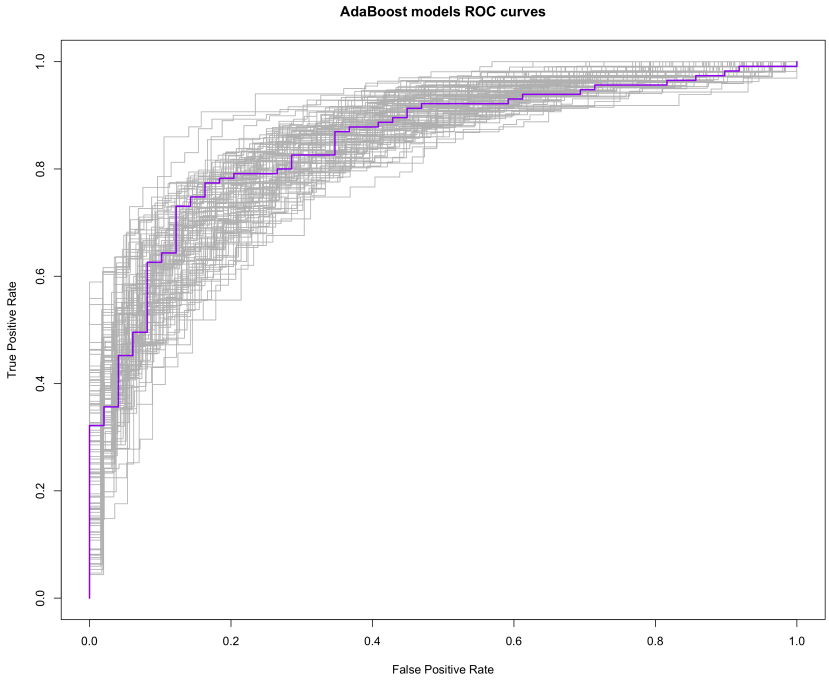

Figure 5.2: ROC Curves for RNA-Seq, 148 Genes

substantia nigra, a region associated with Parkinson's disease. Addressing mitochondrial issues is considered a potential avenue for Parkinson's disease treatment [23, 24].

See our gene connections pattern in Figure 5.3. Our result confirms previous studies on the same pathology. Gene network obtained using the String-DB tool for protein-protein interaction [25]. The genes belonging to the "Mitochondrion" Gene Ontology Cellular Component category are highlighted in red. This particular gene category exhibited statistical significance (p=0.021) in an enrichment analysis conducted using the hypergeometric distribution (Fisher's exact test).

In our initial dataset comprising 22,032 genes, we identified 1,535 mitochondrial genes. While there is a total of 1,665 mitochondrial genes reported in the entire human genome, as per data provided by The Gene Ontology Consortium [26].

Starting from an expression data matrix of 526 genes, we analyzed them using the String-DB tool for functional analysis of gene lists. The top ranking gene category were the Gene Ontology Cellular Compartment and UniProt Keyword "mitochondrion", corresponding to 61 genes. To assess the biological relevance of these mitochondrial genes for classifying Parkinson's disease versus control subjects, we built 100 models using only these 61 genes as predictors. However, the performance was disappointing, with an average AUC (Area Under the Curve) of 0.68 and a maximum of 0.73. This suggests that these genes are relevant in the model, but

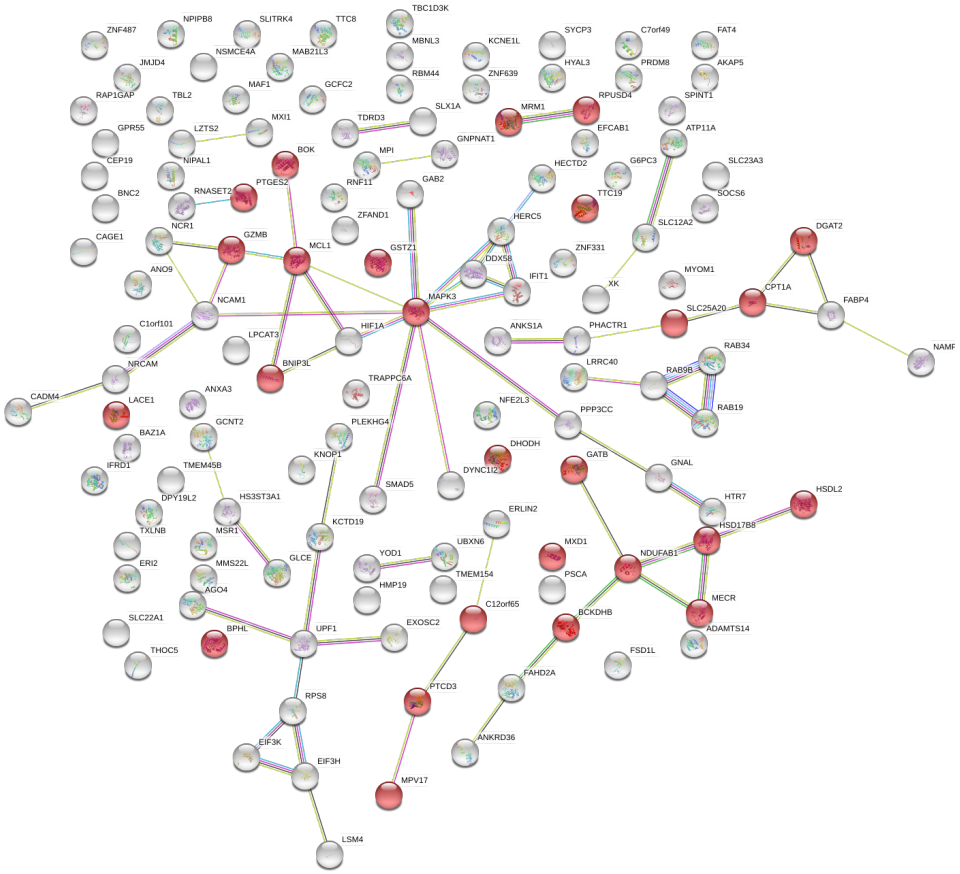

Figure 5.3: Interaction network of 148 predictors genes, obtained by the StringDB tool. The genes belonging to the "Mitochondrion" Gene Ontology Cellular Component category are highlighted in red.

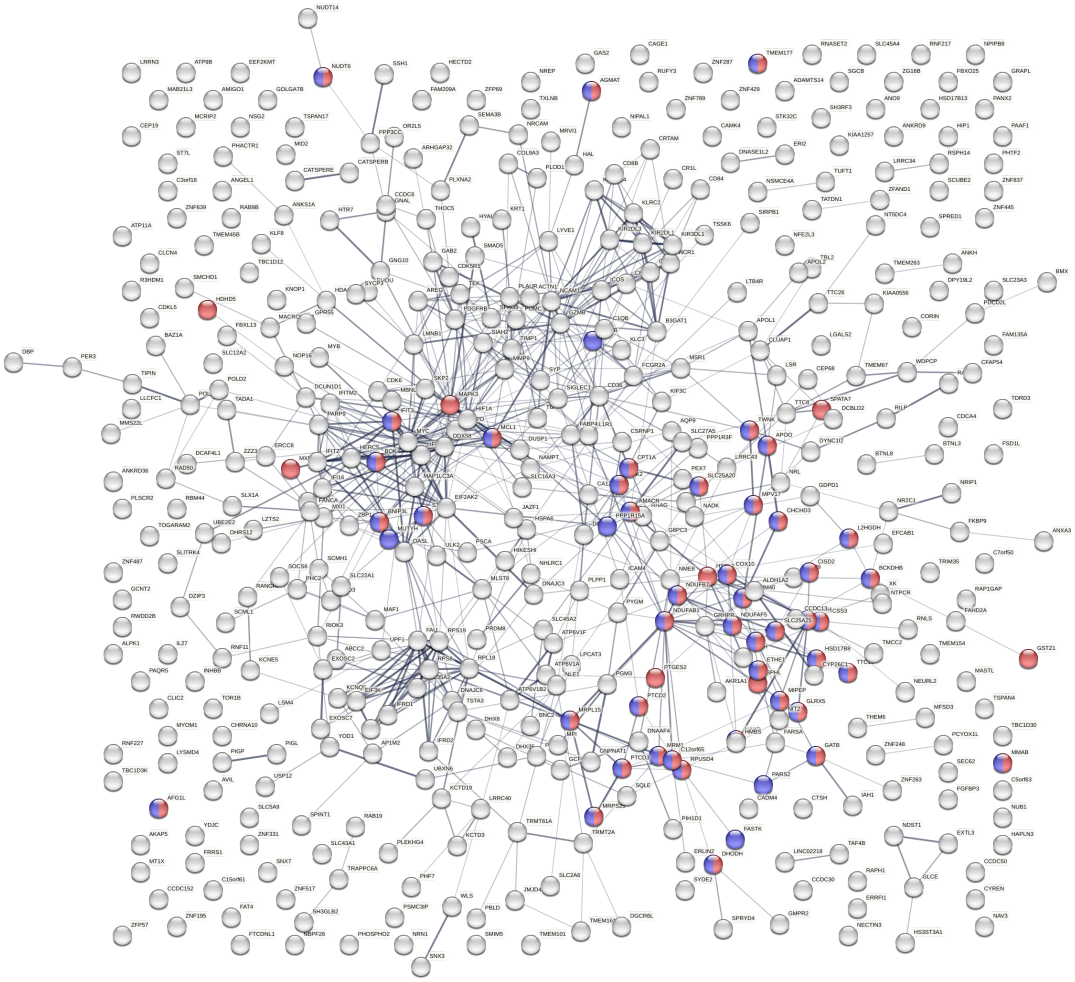

Figure 5.4: Our ML obtained gene network for 526 genes. The genes belonging to the "Mitochondrion" Gene Ontology Cellular Component category are highlighted in red/blue.

only in association to many other predictors not necessarily biologically related to mitochondrial function. Expanding our gene selection to 526 genes results in an increase in the number of mitochondrial genes. However, this expansion does not enhance the performance of our machine learning model. Refer to Figure 5.4 for further details.

These results suggest that while these genes may hold some relevance within the model, their contribution likely hinges on interactions with other predictors, not necessarily related to mitochondrial function.

5.2 Proteomics Result

The outcomes of the application of the ML Adaboost modelling on the proteomics dataset, as outlined in 5.5, are detailed below. It is evident that the model exhibits relevant differences between metrics, with Sensitivity generally smaller than other metrics. In contrast, Specificity consistently demonstrates superior performance.

- **Maximum Mean Accuracy**: Achieved a peak mean accuracy of 0.788, utilizing a set of 75 protein genes.
- **Maximum Mean AUC**: Attained a maximum mean AUC of 0.843, once again leveraging 75 protein genes.
- **Maximum Mean Sensitivity**: Reached a maximum mean Sensitivity of 0.559, utilizing the same 75 protein genes.
- **Maximum Mean Specificity**: Demonstrated a maximum mean Specificity of 0.909, incorporating a set of 135 protein genes.

The accompanying plot illustrates both the absolute maximum and minimum mean values, with and without standard deviation, facilitating a comprehensive comparison.

On the other hand, In our ML analysis with XGBoost and selecting 50% of the Pearson's index correlated proteomics data (2392 proteins) from healthy control and Parkinson's disease subjects, we obtained promising results. The average area under the receiver operating characteristic curve (AUC), a key metric in evaluating the model's discriminatory power, was found to be 0.826 with a relatively low standard deviation of ± 0.059. This indicates a robust ability of the model to distinguish between the two groups based on the proteomic features experimentally measured in CSF samples.

However, it is crucial to delve into the specificities, accuracies, and sensitivities to gain a comprehensive understanding of the model's performance. The average specificity, representing the proportion of true negatives correctly identified, was 0.4175 ± 0.0742. While this value may seem relatively low, it is important to consider the balance between sensitivity and specificity in the context of the specific research question. In the case of Parkinson's disease diagnosis, achieving high sensitivity (0.9032 ± 0.0393) is often prioritized to minimize false negatives, ensuring that individuals with the condition are correctly identified.

The average accuracy of 0.7422 ± 0.0451)indicates the overall correctness of the model's predictions across both classes. The relatively narrow small standard

deviation suggests consistency in the model's performance across different folds of the cross-validation process.

Table 5.1: Performance Metrics for Each Fold and Averaged Results

Fold	**AUC**	**Specificity**	**Accuracy**	**Sensitivity**
1	0.8293	0.5789	0.8393	0.9730
2	0.8919	0.4737	0.7857	0.9459
3	0.7624	0.3684	0.7321	0.9189
4	0.8478	0.4211	0.7500	0.9189
5	0.6711	0.3333	0.6607	0.8158
6	0.7673	0.4444	0.7455	0.8919
7	0.7778	0.4444	0.7455	0.8919
8	0.7447	0.3333	0.7091	0.8919
9	0.7733	0.3333	0.7091	0.8919
10	0.8258	0.4444	0.7455	0.8919
Average	0.8258 ($\pm$0.0587)	0.4175 ($\pm$0.0742)	0.7422 ($\pm$0.0451)	0.9032 ($\pm$0.0393)

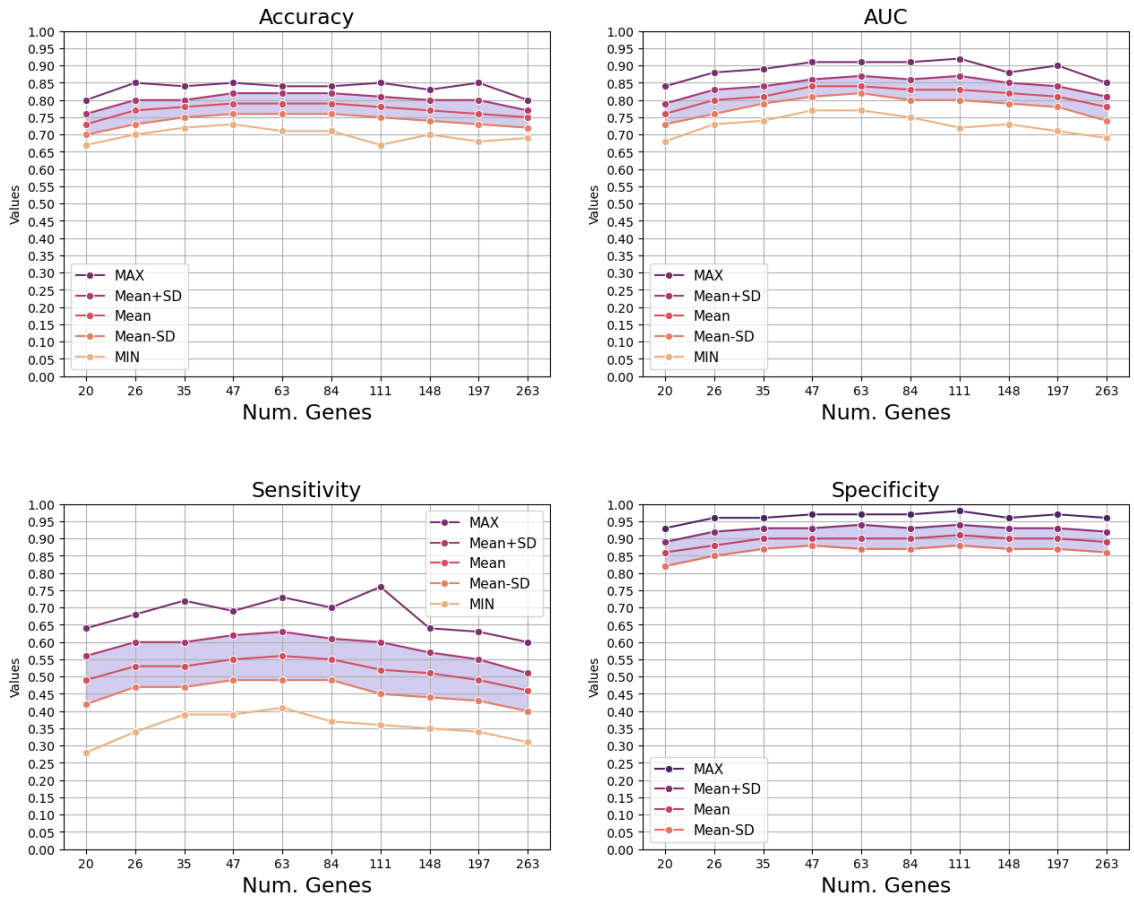

Figure 5.5: Performance Metrics of ML Models on the Proteomics Dataset

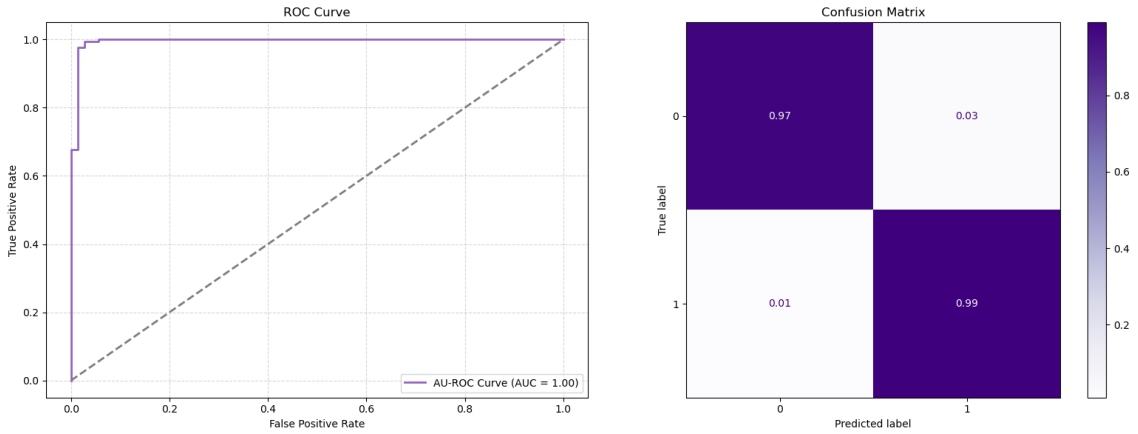

Figure 5.6: UPDRS

5.3 UPDRS and UPSIT Results

Our analysis focused on the motor score, a crucial component of the Unified Parkinson's Disease Rating Scale (UPDRS). We combined data from three UPDRS files, specifically the sum of the totals from columns NP1R, NP1P, NP2P, and NP3. These columns likely represent different aspects of motor function, such as rigidity, bradykinesia, and tremor.

Employing a simple ML algorithm (default XGBoost), we explored the predictive power of the motor score. Apparently, the resulting Area Under the Curve (AUC) approached perfection (AUC = 1.0), that is an overfitting. Such high discriminatory ability suggests that the motor scores are therefore highly correlated with the response variable of the models.

Clinicians routinely rely on motor scores for early detection and monitoring of Parkinsons disease. Our findings, supported by the confusion matrix and ROC curve (see Figure 5.6), reinforce the pivotal role of motor assessment in clinical practice.

In summary, the motor score emerges as a vital tool for identifying and managing PD, emphasizing its clinical relevance and diagnostic accuracy.

We then analyzed the UPSIT data (olfactory scores) from the PPMI archive, focusing on four columns representing different UPSIT booklets. After removing missing values (NaNs), our dataset included 893 patients.

To evaluate the performance of a ML model in predicting UPSIT scores, we employed a XGBoost classifier with the following hyperparameters: learning rate = 0.1, max depth = 1, and n_estimators = 60. We further employed stratified 10-fold stratified cross-validation. In stratified cross-validation, the dataset is divided into folds while preserving the proportion of instances (PD vs HC) for each class. This helps ensure that each fold maintains a representative distribution of classes, which

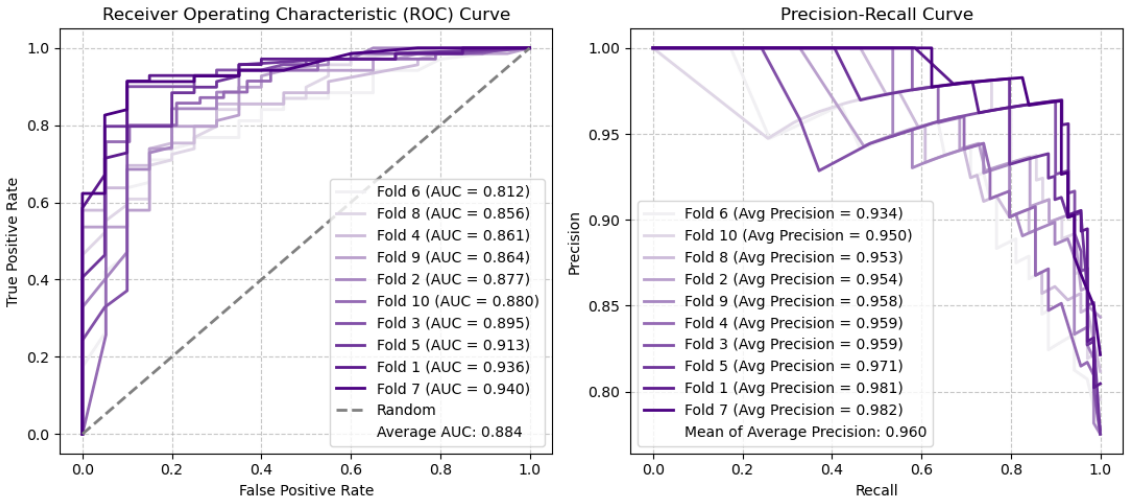

Figure 5.7: ROC and Precision-Recall Curves for UPSIT Score Prediction (10-Fold Cross-Validation)

can be crucial when dealing with imbalanced datasets. we use it to prevent one class from being overrepresented or underrepresented in any particular fold.

The model achieved a promising average Area Under the Curve (AUC) of 0.884, indicating good discrimination between patients with different UPSIT performance. Additionally, the average precision score reached 0.960, demonstrating the model's ability to accurately identify patients with high UPSIT scores (indicating better cognitive function).

For a more detailed visualization of the model's performance across different folds, please refer to Figure 5.7. This figure depicts the ROC curves and precision-recall curves for each fold within the cross-validation process.

Chapter 6

Conclusion and Future Perspective

In this thesis, we have extensively explored transcriptomics, proteomics, and clinical datasets such as UPDRS and UPSIT, utilizing advanced ML techniques to improve predictive outcomes. Despite facing imbalances in our datasets between Parkinson's disease and healthy controls, we discovered unique strengths in each dataset during our analysis. We applied a variety of machine learning algorithms, each showcasing its own advantages. We pinpointed significant genes in RNA-Seq data and important proteins in the proteomics dataset. While we encountered overfitting issues with MDS-UPDRS scores, we also recognized the UPSIT dataset's ability to effectively differentiate between Parkinson's disease and healthy controls.

It is worth noting that several studies, such as the notable investigation by [27], have adopted multimodal machine learning approaches to comprehend Parkinson's disease. These studies integrated proteomic, genetic, and UPSIT test data, yielding impressive results with AUC scores reaching 89.72%. However, when relying solely on transcriptomics data, they achieved a lower AUC of 79.73%, with a sensitivity of 98% and specificity of only 0.12%.

In another study by [28], which exclusively utilized transcriptomics data, they obtained an AUC score of 72%, with a mean sensitivity of 82% and specificity of 47%. In our research, focusing solely on the transcriptomics dataset, we have contributed significantly to this advancement by achieving a noteworthy improvement, with an AUC of 85%, maximum mean specificity of 92%, and sensitivity of 56%.

This review paper [10] examines $n = 110$ papers on how machine learning is used to analyze diverse data from the PPMI, identifying variations in methods and suggesting ways to better utilize the dataset's unique features for biomarker discovery and prognostic prediction.To follow this, our future perspective involves adopting a multimodality approach, aiming to enhance all evaluation metrics in our ML algorithm by combining data such as clinical and *omics datasets.

The promising AUC score attained in our study indicates significant potential. However, there is a recognized need for improvements in sensitivity and specificity, crucial metrics in clinical studies. The algorithm needed to go through a refinement which can target not only selecting the highest AUC score but also an overall boost across all metrics.

Given the expansive nature of RNA-seq data, expected to grow over time, we emphasize the necessity for faster and more robust algorithms. Achieving this requires leveraging the full potential of high-performance computing facilities and implementing parallelism in our algorithms.

The algorithm which we discussed in is not fully automated, we also use the excel. There is a need for automation of our progressive feature selection algorithm.

For Parkinson's diagnosis using the PPMI dataset, we aim to optimize a model that integrates gene expression data with Single Nucleotide Polymorphism (SNP) data. Boosting algorithms with tunable parameters like number of trees and tree complexity (e.g., cp) will be our initial approach. We will explore even more complex models like neural networks, especially for handling the high dimensionality of SNP data (millions of features), potentially employing pruning or filtering to focus on informative features. This combined model has the potential to significantly improve diagnostic accuracy.

In this study we have used PPMI data repository. It is always recommended to reduce the potential overfitting, it is important to of validate evaluations using external datasets like the Parkinson's Disease Biomarkers Program (PDBP) data repository [29].

Despite the strides made, Parkinson's disease data remains relatively scarce. To fully unleash the potential of ML in prediction, increased investments in data acquisition at a global scale are imperative. Combining results from various continents is vital for obtaining more reliable and meaningful predictions, steering clear of potential biases.

Appendix A

Pearson's Correlation

Pearson's correlation coefficient, also known as Pearson's R, is a statistical measure used to quantify the strength and direction of a linear relationship between two continuous variables.

It measures how well the data points fit a straight line (similar to linear regression).

The main formula for Pearson's correlation coefficient (r) is as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual data points for variables X and Y, and $\bar{X}$ and $\bar{Y}$ denote the means of X and Y, respectively.

The value of r ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.

We do not apply Pearson's correlation to UPDRS and UPSIT data as they already have few features for training and test.

Appendix B

Genetic Mutations

To ensure less biased data for ML, we remove specific gene mutations associated with PD. These mutations are as follows:

SNCA stands for alpha-synuclein, which is the gene that encodes for alpha-synuclein protein. Mutations in the SNCA gene have been linked to increased risk of developing synucleinopathies, particularly PD.

Synucleinopathies are a group of neurodegenerative diseases that share the common pathological feature of misfolded alpha-synuclein protein deposits in the brain. The most common synucleinopathy is PD, but other synucleinopathies include Lewy body dementia (LBD), multiple system atrophy (MSA), and pure autonomic failure (PAF) [30, 31, 32].

GBA stands for glucocerebrosidase, which is a lysosomal enzyme that breaks down a complex sugar called glucosylceramide. Glucosylceramide is a major component of the myelin sheath, which insulates and protects nerve cells.

GBA mutations are the most common genetic risk factor for PD. People with GBA mutations have a higher risk of developing PD, and they tend to develop the disease earlier in life than people without GBA mutations.

GBA mutations can also cause other neurological disorders, such as Gaucher's disease and Lewy body dementia (LBD) [33].

LRRK2 stands for leucine-rich repeat kinase 2, which is a protein found in many cells throughout the body, including the brain. LRRK2 plays a role in a variety of cellular processes, including vesicle trafficking, autophagy, and inflammation.

Mutations in the LRRK2 gene are the most common genetic risk factor for PD. People with LRRK2 mutations have a higher risk of developing PD, and they tend to develop the disease earlier in life than people without LRRK2 mutations.

LRRK2 mutations can also cause other neurological disorders, such as Gaucher's disease and Lewy body dementia (LBD) [34].

Bibliography

[1] Andrew Siderowf, Luis Concha-Marambio, Daniel E Lafontant, Chad M Farris, Yuan Ma, Pedro A Urenia, Hien Nguyen, Roy N Alcalay, Lama M Chahine, Tatiana Foroud, Douglas Galasko, Karl Kieburtz, Kalpana Merchant, Brit Mollenhauer, Kathleen L Poston, John Seibyl, Tanya Simuni, Caroline M Tanner, Daniel Weintraub, Aleksandar Videnovic, Seung H Choi, Rebecca Kurth, Chelsea Caspell-Garcia, Christopher S Coffey, Mark Frasier, Leticia M A Oliveira, Samantha J Hutten, Todd Sherer, Kenneth Marek, and Claudio Soto. Assessment of heterogeneity among participants in the parkinson's progression markers initiative cohort using alphasynuclein seed amplification: a crosssectional study. *Lancet Neurol*, 22(5):407–417, May 2023.

[2] Parkinson's progression markers initiative. https://www.ppmi-info.org/. Accessed: Insert-Access-Date-Here.

[3] Zainab Nazari. PPMI Data Analysis Repository. https://github.com/zainabnazari/ppmi/, 2024.

[4] Tanner CM Halliday GM Brundin P Volkmann J Schrag AE Lang AE. Poewe W, Seppi K. Parkinson disease. *Nat Rev Dis Primers*, 23:3:17013, 2017 Mar.

[5] Alexis Elbaz E. Ray Dorsey and et. al. Global, regional, and national burden of parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 17(11):939–953, 2018.

[6] Global burden of disease study (2019) – processed by our world in data. Dataset, 2019. Current number of cases of Parkinson's disease per 100,000 people, in both sexes aged age-standardized.

[7] Zhaolong Ou, Jie Pan, Shaoxuan Tang, Dongxue Duan, Dongmei Yu, Hanqing Nong, and Zhiyong Wang. Global trends in the incidence, prevalence, and years lived with disability of parkinson's disease in 204 countries/territories from 1990 to 2019. *Frontiers in Public Health*, 9, Dec 2021.

[8] Our World in Data. Parkinson's disease prevalence, 1990-2019. Accessed on: Feb 26, 2024.

[9] Charles A Davie. A review of parkinson's disease. *British Medical Bulletin*, 86:109–127, 2008. Epub 2008 Apr 8.

[10] Raphael T. Gerraty, Allison Provost, Lin Li, Erin Wagner, Magali Haas, and Lee Lancashire. Machine learning within the parkinson's progression markers initiative: Review of the current state of affairs. *Frontiers in Aging Neuroscience*, 15, 2023.

[11] Zang-Hee Cho. Review of recent advancement of ultra high field magnetic resonance imaging: from anatomy to tractography. *Investigative Magnetic Resonance Imaging*, 20(3):141–151, 2016.

[12] Julian M. Fearnley and Andrew J. Lees. Ageing and parkinson's disease: Substantia nigra regional selectivity. *Brain*, 114(5):2283–2301, October 1991.

[13] The Parkinson's Foundation. www.parkinson.org. Accessed on Feb 12, 2024.

[14] The Michael J. Fox Foundation. https://www.michaeljfox.org/. Accessed on Feb 12, 2024.

[15] D. Pandey and P. Onkara Perumal. A scoping review on deep learning for next-generation rna-seq data analysis. *Funct Integr Genomics*, 23(2):134, Apr 2023.

[16] Ernesto Picardi. *RNA Bioinformatics*, volume 2284 of *Methods in Molecular Biology (MIMB)*. Springer, 2022.

[17] K. Tsukita, H. Sakamaki-Tsukita, S. Kaiser, L. Zhang, M. Messa, P. Serrano-Fernandez, and R. Takahashi. High-throughput csf proteomics and machine learning to identify proteomic signatures for parkinson disease development and progression. *Neurology*, 101(14):e1434–e1447, Oct 3 2023.

[18] Physiopedia. Cerebrospinal fluid (csf), Year.

[19] PPMI. *PARKINSONS PROGRESSIVE MARKERS INITIATIVE (PPMI) Data User Guid*. PPMI, 2023. https://www.ppmi-info.org/sites/default/files/docs/PPMI

[20] The Aligning Science Across Parkinson's (ASAP). parkinsonsroadmap.org. Accessed on Feb 19, 2024.

[21] Dask Development Team. Dask dataframe documentation, 2024. Accessed on: 2023-2024.

[22] Jason Brownlee. A gentle introduction to xgboost for applied machine learning. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/, 2021. Accessed: [3rd March 2024].

[23] Xiao-Yan Gao, Ting Yang, Yang Gu, and Xiao-Hong Sun. Mitochondrial dysfunction in parkinson's disease: From mechanistic insights to therapy. *Front Aging Neurosci*, 14:885500, June 20 2022.

[24] Martin T. Henrich, Wolfgang H. Oertel, D. James Surmeier, and Fanni F. Geibl. Mitochondrial dysfunction in parkinson's disease – a key disease hallmark with therapeutic potential. *Molecular Neurodegeneration*, 18(1):83, 11 2023.

[25] Protein-protein interaction networks, functional enrichment analysis. Accessed on [2024 Feb.].

[26] The Gene Ontology Consortium. Go: Cellular component - gocc mitochondrion. https://www.gsea-msigdb.org/gsea/msigdb/cards/GOCC_MITOCHONDRION, 2004-2023. Accessed: [29 Feb. 2024].

[27] Mary B Makarious, Hampton L Leonard, Dan Vitale, Hirotaka Iwaki, Lana Sargent, Anant Dadu, Ivo Violich, Elizabeth Hutchins, David Saffo, Sara Bandres-Ciga, Jonggeol Jeff Kim, Yeajin Song, Melina Maleknia, Matt Bookman, Willy Nojopranoto, Roy H Campbell, Sayed Hadi Hashemi, Juan A Botia, John F Carter, David W Craig, Kendall Van Keuren-Jensen, Huw R Morris, John A Hardy, Cornelis Blauwendraat, Andrew B Singleton, Faraz Faghri, and Mike A Nalls. Multi-modality machine learning predicting parkinson's disease. *NPJ Parkinson's disease*, 8(1):35, April 2022.

[28] Ester Pantaleo, Alfonso Monaco, Nicola Amoroso, Angela Lombardi, Loredana Bellantuono, Daniele Urso, Claudio Lo Giudice, Ernesto Picardi, Benedetta Tafuri, Salvatore Nigro, Graziano Pesole, Sabina Tangaro, Giancarlo Logroscino, and Roberto Bellotti. A machine learning approach to parkinson's disease blood transcriptomics. *Genes*, 13(5), 2022.

[29] National Institute of Neurological Disorders and Stroke. Parkinson's disease biomarkers program (pdbp) data repository, 2024. Accessed: 2024.

[30] Aoife P. Kiely, Yasmine T. Asi, Eleanna Kara, Patricia Limousin, Helen Ling, Patrick Lewis, Christos Proukakis, Niall Quinn, Andrew J. Lees, John Hardy, Tamas Revesz, Henry Houlden, and Janice L. Holton. α-synucleinopathy associated with g51d snca mutation: a link between parkinson's disease and multiple system atrophy? *Acta Neuropathologica*, 2013.

[31] Genecards - snca gene. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SNCA. Accessed: March 12, 2024.

[32] Johns Hopkins Medicine. The genetic link to parkinson's disease. https://www.hopkinsmedicine.org/health/conditions-and-diseases/parkinsons-disease/the-genetic-link-to-parkinsons-disease. Accessed: March 12, 2024.

[33] Sophia R. L. Vieira and Anthony H. V. Schapira. Glucocerebrosidase mutations and parkinson disease. *Journal of Neural Transmission*, 129:1105–1117, 2022.

[34] Edoardo Monfrini and Alessio Di Fonzo. Leucine-rich repeat kinase (lrrk2) genetics and parkinson's disease. *Advances in Neurobiology*, 2017.