THE EFFECTS OF REPRODUCIBILITY & REPLICABILITY OF PARKINSON'S DISEASE PROGRESSION PREDICTION USING MACHINE LEARNING

Mohanad Arafe

A thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Software Engineering
Concordia University
Montréal, Québec, Canada

April 2023

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mohanad Arafe**

Entitled: **The effects of reproducibility & replicability of Parkinson's Disease progression prediction using Machine Learning**

and submitted in partial fulfillment of the requirements for the degree of

Master of Software Engineering

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Yang Wang ______________________ Chair
Dr. Marta Kersten-Oertel ______________________ Examiner
Dr. Tristan Glatard ______________________ Supervisor
Dr. Jean-Baptiste Poline ______________________ Co-supervisor

Approved ______________________________
Chair of Department or Graduate Program Director

____________ 20 ______ ______________________________
Dr. Mourad Debbabi, Dean of
Gina Cody School of Engineering and Computer Science

Abstract

The effects of reproducibility & replicability of Parkinson's Disease progression prediction using Machine Learning

Mohanad Arafe

Machine Learning (ML) techniques are growing in popularity for analyzing T1-weighted magnetic resonance imaging (MRI) as it is a promising source of Parkinson's disease (PD) biomarkers. However, there is growing concern within the scientific community regarding the reproducibility of research findings. This reproducibility crisis suggests that a significant proportion of studies may not be reliable which impacts the validity of results in a preclinical setting.

The objective of this paper is to reproduce and replicate the findings of a study by (Shu et al., 2020) that uses ML techniques to predict the progression of PD using conventional MRI and radiomic biomarkers in whole-brain white matter. We aim to assess the reproducibility and replicability of (Shu et al., 2020)'s predictive capabilities using open-source tools. We used the Parkinson's Progression Markers Initiative (PPMI) dataset, the same dataset used by (Shu et al., 2020) and similar analyses to assess the reproducibility of the findings. While we attempted to follow the methods outlined in (Shu et al., 2020) as closely as possible, some details were unclear and we made educated guesses. We introduced variations in the methodological methods, including different cohorts, feature sets, ML algorithms, and evaluation techniques, to assess the replicability of the findings. Our study could not reproduce nor replicate the predictive capabilities of (Shu et al., 2020). The lack of reproducibility and replicability in this paper highlights the importance of adopting open science practices to ensure that proposed biomarkers are robust.

Acknowledgments

I express my sincere gratitude to my supervisor Dr Tristan Glatard and co-supervisor Dr Jean-Baptiste Poline for their support throughout my academic journey. Their expertise, encouragement and mentorship helped me complete this thesis. I am grateful to have learned so much from two of the best researchers in Montreal. I would like to extend my gratitude to the entire LivingPark team, including Mathieu Dugré for his invaluable assistance with the major technical tools utilized in this project. Additionally, I'd like to mention Dr Madeleine Sharp and Dr Yiming Xiao who helped me gain further understanding of Parkinson's disease in a clinical setting. Their knowledge and wisdom helped me understand the importance of the medical context of my thesis.

This work would not exist without the encouragement and motivation from my family, particularly my mother, Lina Kudsi, and my eldest brother, Feras Arafe. Their belief in me was the driving force behind my decision to pursue a Master's degree immediately after getting my Bachelor's degree. I am deeply grateful for their support and for helping me to push my limits. Finally, I'd like to thank a very special person, my fiancée Khadija, her unwavering support and belief in my abilities made a big impact. I'd like to thank her for her sacrifices and support in making this thesis possible.

Contribution of Authors

This paper represents a collaborative effort, and as such, it includes several co-authors who played crucial roles in making this work possible. We are currently in the process of publishing our paper on bioRxiv and aiming to submit it to PLOS One. Here are the contributions of each authors.

Mohanad Arafe was responsible for the entire software development and writing the thesis draft.

Tristan Glatard, Jean-Baptiste Poline were the supervisor and co-supervisor's of the project. They help me conceptualize the entire work.

Yohan Chatelain, Mathieu Dugré, Andrzej Sokolowski, Michelle Wang built and tested the PPMI data retrieval pipeline, and provided technical support.

Nikhil Bhagwat, Yiming Xiao, Madeleine Sharp helped interpret clinical data, and provided insights into the diagnosis and treatment of PD.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disease characterised by both motor and nonmotor features. Although there is currently no known cure for PD, correctly identifying the disease in its early stages is crucial for providing appropriate initial treatment. One promising area for detecting PD is through the use of neuroimaging techniques, such as magnetic resonance imaging (MRI). Machine Learning (ML) techniques are growing in popularity for the examination of T1-weighted MRI scans. ML can help researchers identify pattern changes in the brain that may be early indicators of PD. However, the usage of ML in PD research has raised important questions about the reproducibility and validity of findings.

In this chapter, we will focus on the reproducibility and replicability of MRI-derived PD biomarkers. We will begin by defining key terms related to reproducibility and replicability and discuss the current state of the reproducibility crisis. We will then examine the current usage of ML in neuroimaging research. Finally, we will outline the objectives of this thesis.

1.1 Reproducibility Definitions

The terms "reproducibility" and "replicability" are distinct and may cause confusion. In fact, the first definition dates back to the early 1990s, when geophysicist Jon Claerbout drew attention to the issue [10]. He defined reproducibility as "running the same software on the same input data and obtaining the same results" and replicability as "obtaining sufficiently similar results by designing and running new code based on a published description of a model" [25]. This started a trend of scientists and researchers voicing their opinions on the proper usage of the terms. For instance, in social sciences, Harvard professor Gary

King uses "replication" to cover all related concepts to the term [22]. The Association for Computing Machinery refers to reproducibility as "Different team, same experimental setup" and replicability as "Different team, different experimental setup" [3]. Needless to say, the wide variety of definitions presents a challenge in standardizing the usage of the terms. The work in [5] conducted a review of the usage of the terms and grouped findings in three categories:

(A) The terms are used with no distinction between them.

(B1) "Reproducibility" refers to instances in which the original researcher's data and computer codes are used to regenerate the results, while "replicability" refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.

(B2) "Reproducibility" refers to independent researchers arriving at the same results using their own data and methods, while "replicability" refers to a different team arriving at the same results using the original author's artifacts.

A	**B1**	**B2**
political science	signal processing	microbiology
economics	scientific computing	computer science
	econometry	
	epidemiology	
	clinical studies	
	internal medicine	
	physiology (neuro)	
	computational biology	
	biomedical research	
	statistics	

Table 1: Grouping of terminologies by discipline from [5]

Table 1 illustrates the distribution of the usage of the terms among different scientific disciplines. The results show that the B1 definition is widely utilized. Therefore, in this work, we follow B1's defintions of reproducibility and replicability. Regardless of the various definitions and disciplines that exist, these terms simply aim to describe the attempt to assess the validity of results being produced. This leads us to our next topic of conversation, the reproducibility crisis.

1.2 Reproducibility Crisis

Scientists today have access to a wide variety of software, computing and communication tools that make research easier than ever. In the past, researchers relied on limited resources, such as paper and pencil to perform their experiments. Today, scientists have the ability to tackle more complex questions and share their findings to the world. Figure 1 depicts the number of computer science publications released per year over the past 50 years. The figure demonstrates the significant increase in the yearly release of papers, reaching a peak of 400,000 releases per year. In a survey conducted between the years 2000 and 2016, the work in [20] found that more than 9,000 individuals published at least 72 "full" papers per year. That is the equivalent of one paper every 5 days.

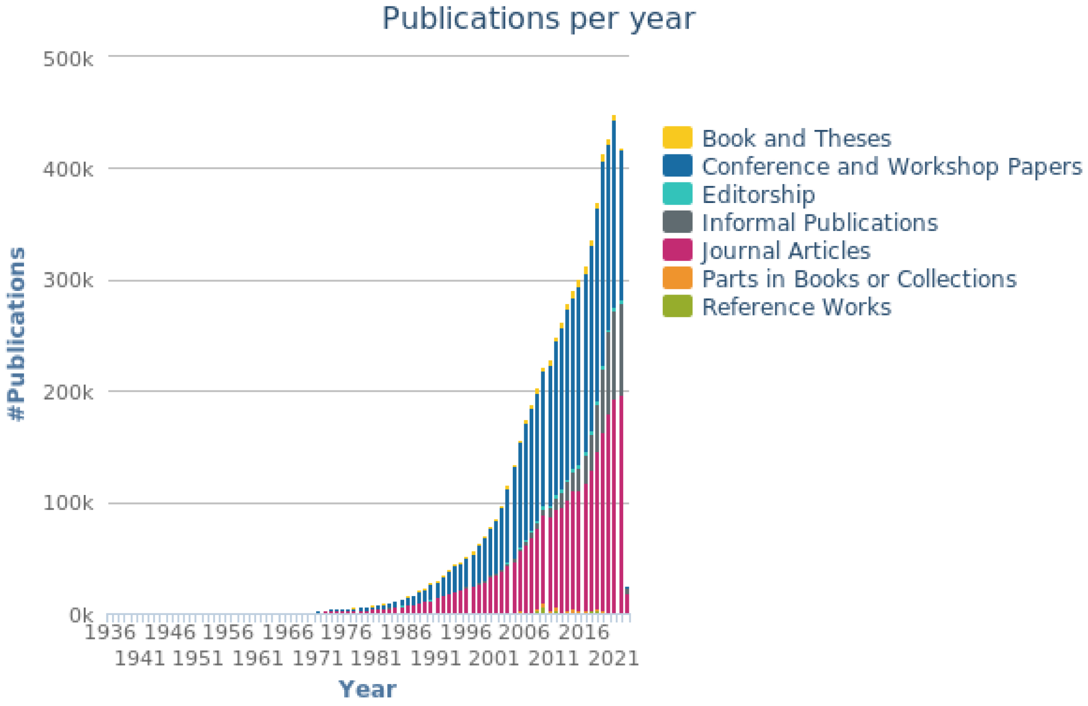

Figure 1: Number of computer science publications released per year. [15]

The scientific community relies on the validity of published results to make advancements in major fields. However, with limited resources and incentives, there is a lack of motivation for researchers to reproduce published results. The dialog of replication studies gained more interest in 2015 due to the Open Science Collaboration's Reproducibility Project. The project aimed to assess the reproducibility of psychological science by having 270 researchers across 41 institutions replicate 100 published studies. The results showed that only 39 of the 100 studies were successfully replicated. The term "Reproducibility crisis" refers to the "large and growing proportion of studies published across disciplines [that] are unreliable due to the declining quality and integrity of research and publication practices"[13]. In 2016, the work in [4] conducted a survey in which 1,576 researchers gave their views on reproducibility in research. A staggering 70% have said that they tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Additionally, 90% agreed that there is a slight or significant crisis on-going.

One of the factors contributing to the on-going crisis is the lack of a standardized protocol to report results. Firstly, researchers frequently do not publish their code. Although having the code does not guarantee reproducibility, as Nicolas Rougier at France's National Institute for Research in Computer Science and Automation in Bordeaux has argued [19], it certainly provides a foundation for gaining deeper understanding of a published study. Second, 60% of the participants in [4]'s survey have agreed that the pressure to publish and selective reporting are two of the main factors that lead to reproducibility issues. One of the contributing factors to academic success is publishing good papers that earn reputable recognition [29]. Researchers who aim to advance their careers are less likely to work on replication studies than publishing original work [1]. For example, [26] said that replication studies are "low-prestige, mundane, unoriginal, or non-academic and therefore not encouraged by faculties". Due to the competitive nature of research, the pressure to publish often results in negligent reporting of findings. Another contributing factor to the crisis is the publication bias towards publishing positive results. The work in [31] argues that it is easier for researchers to get their papers released in prestigious journals when sharing positive results.

On the other side of the spectrum, there are those who believe we are not necessarily in a crisis. The work in [16] argues that the unrealistic expectations of individual study reproducibility may alter the perception of the crisis. Additionally, [16] makes an interesting point by questioning if replication studies themselves may have reproducibility issues. As discussed in the previous section, some researchers have different definitions of reproducibility and what constitutes as a successful replication. The phenomenon of the reproducibility crisis is an on-going debate that will probably continue for the next decades to come until a feasible solution is standardized.

1.3 Neuroimaging analysis with Machine Learning

The mapping of brain function has been a goal of scientists for centuries. In the late 20th century, the invention of noninvasive methods such as MRI has advanced the field known today as neuroimaging. In essence, neuroimaging refers to the ability to study and vizualize the structure and function of the brain. T1-weighted MRI is a promising source of biomarkers as it can provide detailed information about brain structure [24]. This method can be used to compare the brain structure of a healthy individual and one who is suffering of a disease. According to the World Health Organization, the death rate caused from PD is increasing faster than any neurological disease, and the prevalence of PD has doubled in the past 25 years [2]. PD was first described in 1817 by Dr. James Parkinson as a “shaking palsy”. It is a chronic, progressive neurodegenerative disease characterised by both motor and nonmotor features. Common motor symptom include resting tremor, bradykinesia, and muscular rigidity [12] while common non-motor symptoms include cognitive, neuropsychiatric, sleep, autonomic and sensory disturbances [28]. The Hoehn and Yahr scale (HYS) is a clinical scale that describes the severity of PD. The scale consists of 5 stages, ranging from Stage 1 (mild) to Stage 5 (severe). Table 2 describes the common symptoms at each stage.

HYS	Symptoms
1	Only unilateral involvement, usually with minimal or no functional disability
2	Bilateral or midline involvement without impairment of balance
3	Bilateral disease: mild to moderate disability with impaired postural reflexes; physically independent
4	Severely disabling disease; still able to walk or stand unassisted
5	Confinement to bed or wheelchair unless aided

Table 2: Summary of PD symptoms per HYS from [6]

There exists several neuroimaging analysis techniques that can be applied to T1-weighted MRI images to gain insights into brain structure of individuals with a disease such as PD. These methods include voxel-based morphometry, cortical thickness measurements, and deformation-based morphometry. Typically, open source software packages such as FreeSurfer [14] or FSL (FMRIB Software Library) [21] are used for conducting these analyses. During the pre-processing step, these software packages are used to extract brain tissues, such as the white matter (WM). In the statistical assessment phase, the measures obtained from the pre-processing step are evaluated to gain a deeper understanding of the data and interpret the results.

In recent years, ML has emerged as a powerful tool for neuroimaging analysis due to its ability to identify complex patterns and relationships. Researchers have used ML algorithms to perform disease-associated predictions based on MRI data. For example, several papers have used ML techniques to make predictions about PD based on MRI data [8] [33]. ML has also been used to predict disease progression by analyzing MRI images taken over time [32]. Despite its potential, MRI-based measurements of PD have yet to be widely adopted in clinical and research settings, in part due to the lack of reliability, robustness, and reproducibility of such measures. The lack of reproducibility of MRI measures of PD likely originates in variability in population sampling, image acquisition parameters, and image analysis conditions. For instance, the work in [37] showed that longitudinal measures of cortical thickness revealed conflicting results for PD progression. Moreover, the work in [18] has shown that measurements of anatomical volume and cortical thickness are affected by the version of FreeSurfer, workstation type, and version of operating system used. Therefore, the variability observed across imaging studies significantly threatens the reliability of MRI-derived biomarkers.

1.4 LivingPark initiative

The LivingPark initiative investigates whether existing potential MRI biomarkers of PD are impacted by analytical and dataset variability. It also examines whether this variability can be leveraged, using ML or statistical methods to improve their quality. The initiative investigates the following questions:

- What is the impact of the image analysis software toolbox on the reliability of potential biomarkers?
- What is the replicability of potential biomarkers across different datasets?
- Can we improve biomarkers by combining measurements from multiple toolboxes or datasets?

In order to address the research questions, the LivingPark initiative intends to replicate eleven MRI-derived measures of neurodegeneration listed in [24], the most recent comprehensive review of potential MRI biomarkers in PD. The repository containing all the replication studies can be found here (https://github.com/LivingPark-MRI/). The initiative's aim is not to debunk previous findings, but rather to use the analytical techniques in these studies as a basis for comparing different analysis choices and measuring their impact.

1.5 Thesis objectives

The objective of this thesis is to examine the impact of reproducibility and replicability of MRI-based biomarkers for PD progression prediction using ML. We performed a replication study of a previously published paper that contributes to the LivingPark initiative. We conducted a reproducibility experiment to evaluate the ability to reproduce the original findings. Additionally, we conducted a replicability experiment that introduces methodological changes to the original pipeline and examined the robustness of ML to these changes.

Chapter 2

Background

In this chapter, we will discuss current research on the use of MRI-derived measurements in PD using ML. To achieve this, we conducted a literature review of recent studies in the field. Through our review, we identified a paper using inclusion and exclusion criteria that will serve as the primary focus of our reproduction & replication experiment.

2.1 Selecting a paper for reproduction & replication

To select a paper for our reproduction and replication experiment, we conducted a thorough literature review to carefully choose the study for reproduction and replication. We used PubMed to extract a list of publications based on MRI-derived biomarkers for PD. Following that, we only selected the papers that used ML techniques. For each paper, we recorded the main objectives, class being predicted (e.g. PD versus Healthy Control), dataset, imaging protocols, features, software tools, ML algorithm, and performance metrics. We only considered papers that used data from the Parkinson's Progress Markers Initiative (PPMI) database. The PPMI is a landmark study that launched in 2010 with a mission to identify biomarkers of PD [23]. The study provides data on PD patients, including demographic information, clinical assessments, imaging data, and biological samples. The PPMI dataset is a valuable resource for the PD research community, and it has been utilised in many studies to advance the understanding of PD. We filtered out papers by applying exclusion criteria. The criteria includes (i) absence of MRI-based prediction of or association with PD-related phenotype (ii) use of functional or diffusion MRI (iii) use of non-publicly available software toolbox (iv) sample population less than 30 (v) use of non-MRI imaging (vi) performs prediction of PD v.s. Healthy Control. After reviewing the literature and narrowing down the options, we selected the study *Predicting the progression of Parkinson's disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter* by [34] as the focus of our reproduction and replication. The full literature review is available in Appendix A.

The work in [34] developed and validated a radiomics signature using whole-brain WM and clinical features to predict the progression of PD over 3 years. They segmented WM masks from T1-weighted MRI scans, extracted radiomic features from the WM, trained a classifier with these features and evaluated its capability to predict PD progression. Radiomics is a field of research aiming to extract high-dimensional data from clinical images. Radiomic features can be classified into four types, namely shape features, first-order statistics features, second-order statistics features, and higher-order statistics features, which are obtained using different methods and provide various kinds of information about the images [38]. The study used the HYS to track the severity of PD over time. Patients were classified into the disease progression group if the HYS score increased by any number over the 3-year follow-up, or into the stable disease group if the HYS score did not increase. Seventy-two patients were included in the progression group and seventy-two patients were included in the stable group. [34] used SPM12 to extract whole-brain WM masks from which 378 radiomic features were extracted using the A.K software (Quantitative Analysis Kit, version 1.2, GE Healthcare). The maximum relevance minimum redundancy (mRMR) algorithm reduced the dimensionality of the extracted features to 7 features. Finally, they trained a support vector machine (SVM) with a linear kernel to construct the radiomics signature. The radiomics signature achieved an Area under the ROC Curve (AUC) of 0.795. The pipeline used in [34] is depicted in Figure 2. [34] also developed a joint model that combined both the radiomics signature and clinical features extracted from PPMI into a single model. This model demonstrated a slightly better performance with an AUC of 0.836. However, our study focuses primarily on evaluating the reproducibility and replicability of imaging biomarkers, hence, we solely focused on the radiomics signature in our study. We chose [34] as the focus of our reproduction and replication since it highlights the potential of using ML techniques to develop biomarkers for disease progression. Identifying biomarkers that can predict the progression of PD is essential to support the development of new therapies and track reponses to these new therapies.

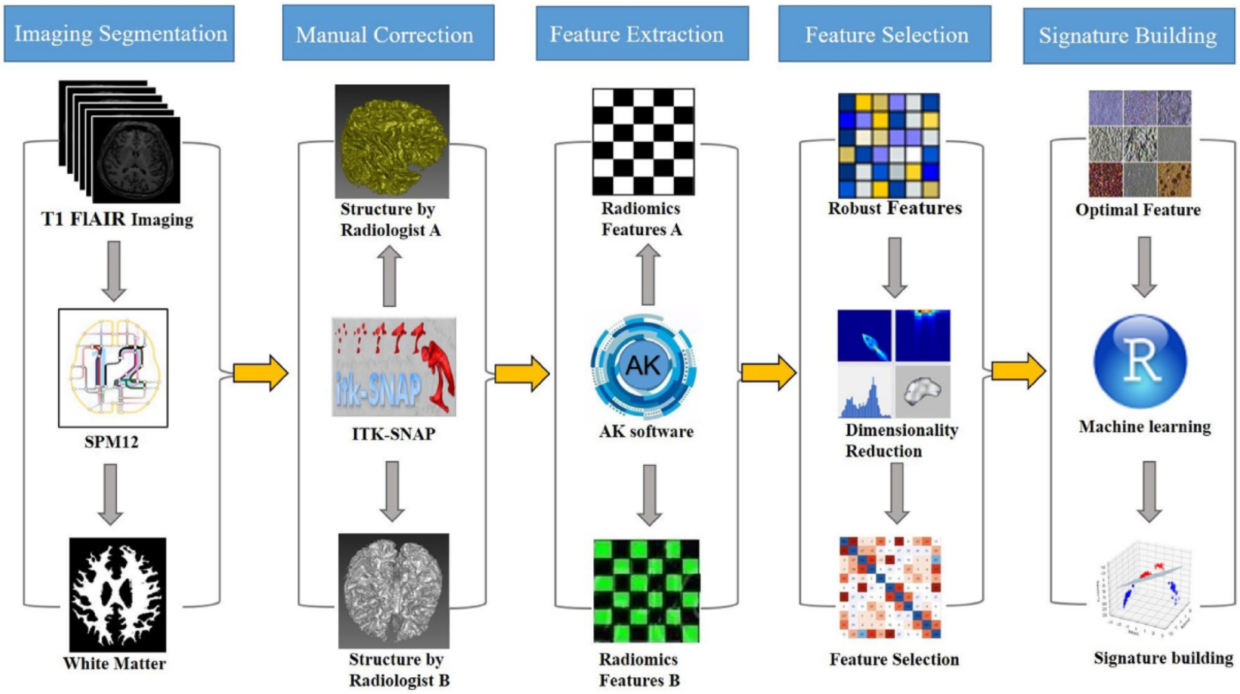

Figure 2: The pipeline was developed and extracted from [34] and consists of five main steps: WM segmentations with SPM12, manual corrections by two experienced neurologists using ITK-Snap, extraction of radiomic features using the A.K Software, reduction of features using the mRMR algorithm, and model building.

2.2 Review of ML based MRI-derived measures

In our literature review, we came across different ML techniques used to identify potential biomarkers of PD using MRI data. For instance, the work in [7] used deep learning (DL) techniques to classify PD and healthy control (HC) patients using 3D Convolutional Neural Networks (CNN), achieving high AUC values of 0.98. There have also been other studies that used unsupervised learning algorithms to classify PD, HC, and scans without evidence for dopaminergic deficit (SWEDD) patients. The work in [35] combined Kohonen self-organizing map (KSOM) and a least squares support vector machine (LS-SVM) to extract features from brain MRI's and classify subjects reporting accuracy values up to 99%. In a review of 110 papers that used ML for the prediction of PD using PPMI data, the work in [17] reported that almost 87% of papers made use of clinical data for their research. As for neuroimaging markers, 79 studies used some form of brain imaging data such as structural MRI, diffusion tensor imaging data, and dopamine transporter scans. The variety of MRI preprocessing and ML techniques observed in [17] suggests that MRI preprocessing steps should be validated and standardized moving forward.

One paper that caught our attention was called *Automated Categorization of Parkinsonian Syndromes Using Magnetic Resonance Imaging in a Clinical Setting* by [9] [9]. The authors categorised various parkinsonian syndromes by segmenting MRI scans using FreeSurfer. They measured brain volumes from 13 segmented regions of interest (ROI) and used them as input features to their ML models. Furthermore, they validated their results by using two independent cohorts from two separate studies. They achieved AUC values ranging from 0.839 to 0.871. The paper attracted our attention as it used recognized tools and techniques and published trustworthy results. We did not choose this study as the focus of our reproduction and replication as it did not use PPMI data, which is an important methodological aspect of our research. However, we considered the features used in [9]'s study as an interesting variation we could introduce to [34]'s features.

2.3 Conclusion of Literature Review

This paper is an attempt to both reproduce and replicate the study "*Predicting the progression of Parkinson's disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter*" [34]. Our study first assesses the reproducibility of the findings in [34] using the same dataset and similar analysis software as in the original study. Further, we assess the replicability of the findings by introducing variations in the methodological methods, including different cohorts, feature sets, ML algorithms, and evaluation techniques.

Chapter 3

Predicting Parkinson's disease progression using MRI-based white matter radiomic biomarker and machine learning : a reproducibility and replication study

Mohanad Arafe[1], Nikhil Baghwat[2], Yohan Chatelain[1], Mathieu Dugré[1], Andrzej Sokolowski[1], Michelle Wang[2], Yiming Xiao[1], Madeleine Sharp[2], Jean-Baptise Poline[2], Tristan Glatard[1]

Pending publishment in: bioRxiv

[1]Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
[2]Department of Neurology and Neurosurgery, McGill University, Montreal, Canada

[34] reported an AUC of 0.795 and a relative standard deviation (RSD) of 3.23 using a linear SVM trained from radiomic features. Our objective was to assess the reproducibility and replicability of the original study. For the reproducibility experiment, we reproduced the cohort, feature set, model, and evaluation technique that resembled closest to [34]'s methods. We then compared the AUC and RSD of our reproducibility experiment to the results reported in [34]. For the replicability experiment, we created various replications of [34]'s study by creating several cohorts using PPMI data, various feature sets, and evaluation techniques. We tested each possible combination of these and compared the resulting AUC values to those achieved in (Shu et al,. 2020), as well as all tested configurations to assess the stability of [34]'s results to methodological perturbations. Figure 3 shows a summary of all the configurations tested.

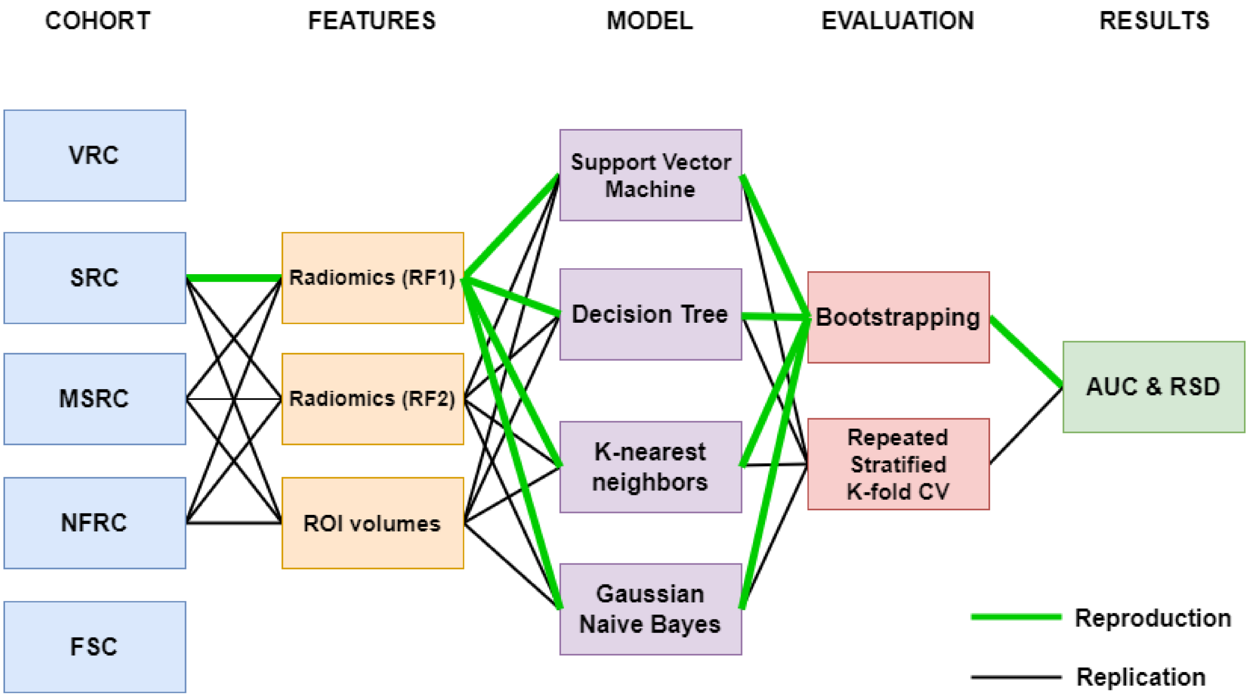

Figure 3: Combination of all configurations tested. The cohorts include the Verio Reproduction Cohort (VRC), Siemens Replication Cohort (SRC), Multiple Scanner Replication Cohort (MSRC), No Filter Replication Cohort (NFRC) and the Functional State Cohort (FSC). The features include the mapped radiomic features (RF1), selected radiomic features (RF2) and volumes of regions of interest (ROI volumes). The green line represents the configuration for the reproducibility experiment while the black lines are the different configurations of the replicability experiment.

3.1 Cohort construction

The [34] study used data from the PPMI database. Their study population was constructed by including PD patients that were matched by age, sex, and baseline HYS score across each group. The MRI data was collected from 32 international sites using a Siemens Verio 3T MRI machine. The protocol used for data acquisition was standardized by PPMI protocols and included the following parameters: repetition time = 2300 ms, echo time = 2.98 ms, inversion time = 900 ms, slice thickness = 1 mm, field of view = 256 mm, and matrix size = 240 × 256. Each patient was evaluated over the course of 3 years. Patients with HYS scores higher at follow-up than at baseline were included in the progressive set (n=72) and patients with the same HYS score at follow-up and at baseline were included in the stable set (n=72).

In our reproducibility experiment, our objective was to build a cohort that is as close as possible to the one in [34]. We filtered the PPMI database for patients meeting the following inclusion criterias:

- **C1**: received a diagnosis of idiopathic PD;
- **C2**: has a pair of visits spaced 3 years apart, with a T1-weighted MRI available at the first visit;
- **C3**: has HYS data available at both visits.

After constructing the cohorts, we conducted sanity checks which ensured that (i) both groups (stable and progressive) are equal size (ii) no patient is present more than once in each group and (iii) patients in one group aren't in the other group and vice versa. We generated two groups of cohorts which will be referred to as the **reproduction cohort** and the **replication cohorts**.

3.1.1 Reproduction cohort

PPMI data was accessed on November 22$^{\text{nd}}$, 2022. For the reproduction cohort, we performed an advanced image search in PPMI with the following filters:

- Research Group: PD
- Acquisition Type: 3D
- Field Strength: 3T
- Slice Thickness: 1 mm
- Manufacturer: Siemens

- Manufacturer Model: Verio
- Weighting: T1

We will refer to this cohort in the paper as the Verio Reproduction Cohort (VRC). This cohort uses all the filters that we could extract from [34]'s methods section.

3.1.2 Replication cohorts

In case the VRC does not successfully reproduce [34]'s cohort, we constructed 3 sub-cohorts with increasingly permissive filters. For the first cohort, we included patients scanned using a Siemens manufactured MRI machine. We performed an advanced image search in PPMI with the following filters:

- Research Group: PD
- Acquisition Type: 3D
- Field Strength: 3T
- Slice Thickness: 1 mm
- Manufacturer: Siemens
- Weighting: T1

We will refer to this cohort in the paper as the Siemens Replication Cohort (SRC). The VRC only includes patients that were strictly scanned using a Siemens Verio MRI machine. The manufacturer model filter is slightly more permissive in the SRC than in the VRC, which is meant to accommodate variations in manufacturer model descriptions throughout the PPMI study.

In case the SRC does not successfully replicate [34]'s cohort, we constructed two more cohorts with increasingly permissive filters. The Multiple Scanner Replication Cohort (MSRC) includes patients scanned with any manufactured MRI machine and was obtained using the following filters:

- Research Group: PD
- Acquisition Type: 3D
- Field Strength: 3T
- Slice Thickness: 1 mm
- Weighting: T1

The No Filter Replication Cohort (NFRC) includes patients with an MRI of any field strength and a slice thickness between 1 mm and 1.2 mm. We used the following filters:

- Research Group: PD
- Acquisition Type: 3D
- Slice Thickness: 1 mm $\leq$ 1.2 mm
- Weighting: T1

In addition to the cohorts described, we will also construct a functional state cohort (FSC) based on the functional state of patients at each visit. The PPMI protocol requires that clinical assessments be conducted twice per visit in different functional states ("ON state" vs "OFF state"). The functional state of a patient refers to the patient's response to medication which importantly affects clinical measures such as the HYS and should therefore be taken into account when building the cohort. [34] did not mention if patients in their cohort were in the ON/OFF state. The variables related to a patient's functional state during a visit are reported in MDS-UPDRS Part III evaluations and include:

- PDSTATE (ON/OFF): the current functional state of the patient
- PDTRTMNT (0/1): 1 if the participant is on PD medication or receives deep brain stimulation, 0 otherwise
- PDMEDTM: time of most recent PD medication dose
- PDMEDDT: date of most recent PD medication dose

In this FSC, we modified inclusion criterion **C3** so that HYS measures of a given patient were obtained with the same PDSTATE (ON or OFF) at both visits. This is meant to ensure that HYS measures were consistently obtained between visits and are therefore comparable. Moreover, the MDS-UPDRS Part III evaluations available in PPMI contain inconsistencies and missing data that we corrected as described in Appendix B & C. We used the following imaging filters on the resulting data:

- Research Group: PD
- Acquisition Type: 3D
- Field Strength: 3T
- Slice Thickness: 1 mm
- Weighting: T1

For each sub-cohort, we used the list of patients returned by the PPMI query and kept those that have a pair of MRI visits spaced 3 years apart. Furthermore, we matched patients from both groups based on age, sex and baseline HYS.

	VRC	**SRC**	**MSRC**	**NFRC**	**FSC**
Research Group	PD	PD	PD	PD	PD
Acquisition Type	3D	3D	3D	3D	3D
Field Strength	3T	3T	3T	any	3T
Slice Thickness	1mm	1mm	1mm	1mm $\leq$ 1.2mm	1mm
Manufacturer	Siemens	Siemens	any	any	Siemens
Manufacturer model	Verio	any	any	any	any
Weighting	T1	T1	T1	T1	T1

Table 3: Summary of PPMI filters used in each cohort.

3.2 MRI Feature extraction

We extracted two sets of image features for each cohort. The first set of features (**F1**) is radiomics-based as per [34]'s methods. The second set of features (**F2**) consists of WM, gray matter (GM) and ventricle volumes measured from known ROI's involved in parkinsonian syndromes [9]. We chose to implement the F2 features because [9] achieved high performances (AUC 0.839 to 0.871) in their study. Moreover, some of the ROI's are in the basal ganglia, which is strongly associated with PD. [9] used two independent cohorts to validate their results, which increases the reliability and trustworthiness of these features.

3.2.1 Segmentation of T1-weighted images

We used two different software tools to perform automatic whole-brain segmentation for **F1** and **F2**. For F1, we used the Segmentation module of SPM12 with default parameters to get the tissue probability masks and build a WM binary mask for each patient. For F2, we used FreeSurfer v6.0 to get the ROI volumes needed.

3.2.2 Quality Control

In [34], two experienced neuro-radiologists used ITK-snap to manually modify WM volumes. The modifications include (i) removal of nonbrain tissue, brainstem and cerebellum and (ii) correcting segmentation errors in WM tissues. We used 3D Slicer v.5.0.3 to visualize and assess the quality of WM segmentations produced by SPM12. For each MRI scan, we reviewed the axial, coronal and sagittal slices. We used the following QC failure criteria:

- There is WM outside of the segmented WM mask;
- There is GM inside the segmented WM mask;
- The MRI has any common artifacts;
- The MRI has a low signal-to-noise (SNR) ratio;

We used FreeSurfer's QA tools [3] to assess the quality of the segmentations. The tool can be used for outlier detection, SNR calculation, WM intensity measurement and collecting detailed volume snapshots.

3.2.3 Radiomic features

The A.K. software (Artificial-Intelligent Radio-Genomics Kits; GE Healthcare, Chicago, IL, USA) used in [34] is not publicly available. Therefore, we used PyRadiomics [36], an open-source Python package for the extraction of radiomics features. PyRadiomics can extract a total of 56 features relevant to our study, including 24 gray level cooccurrence matrix (GLCM) features, 16 gray level size zone matrix (GLSZM) features and 16 gray level run length matrix (GLRLM).

[34] extracted a total of 378 features, including 42 histograms features, 10 Haralick features, 9 FormFactor features, 126 GLCM features, 180 GLRLM features, and 11 gray level region matrix features (GLZSM). From these 378 features, the authors used the maximum relevance minimum redundancy (mRMR) algorithm to extract the following top 7 features to train the model:

- Feature 1: GLCMEntropy_AllDirection_offset1
- Feature 2: RunLengthNonuniformity_angle45_offset7

[3] https://surfer.nmr.mgh.harvard.edu/fswiki/QATools

- Feature 3: Correlation_angle45_offset1
- Feature 4: HaralickCorrelation_angle90_offset4
- Feature 5: ShortRunEmphasis_angle0_offset7
- Feature 6: HaralickCorrelation_AllDirection_offset7
- Feature 7: Inertia_AllDirection_offset4

We extracted two sets of features using PyRadiomics. The first set, **RF1**, includes 5 PyRadiomics features that best match the 7 A.K software features from [34], namely:

- Feature 1: Joint Entropy
- Feature 2: Run Length Non Uniformity
- Feature 3 / Feature 4 / Feature 6: Correlation
- Feature 5: Short Run Low Gray Level Emphasis
- Feature 7: Contrast

The mapping between A.K software and PyRadiomics features is not exact. Indeed, the A.K software, unlike PyRadiomics, provides every feature at a specific angle and offset. In PyRadiomics, for each feature class, the value of a feature is calculated for each angle separately, after which the mean of these values is returned. The exact definitions of these features are available in the PyRadiomics documentation[4] and in the supplementary material of [34], Table S2. [34] used the mRMRe R package [11] to identify the top 7 radiomic features. As such, we will use the mRMRe package with R v4.2.1 on the PyRadiomics features extracted per cohort with K=7 in the second set, **RF2**.

3.2.4 Volumes of Regions of Interest

We used FreeSurfer to extract 13 ROI's that contribute to parkinsonian syndromes as shown in [9]. Those include the midbrain, pons, putamen, posterior putamen, caudate, thalamus, pallidum, precentral cortex and insular cortex in the gray matter, the superior cerebellar peduncle, and the cerebellum white matter including the middle cerebellar peduncles in the white matter and the third and fourth ventricles. Every region's volume was used as input features in our ML algorithms.

[4] https://pyradiomics.readthedocs.io/en/latest/features.html

3.2.5 Feature Normalization

We normalized and centered the features using scikit-learn's StandardScaler, resulting in the following transformation:

$$Z = \frac{x - \mu}{\sigma}$$

where μ is the mean of the training samples and σ is the feature standard deviation in the training sample. We then applied this transformation to the test set, reusing the mean and standard deviation values learned in the training set.

3.3 Machine Learning Model

To predict disease progression, [34] trained a linear SVM based on the 7 features extracted and selected from segmented WM masks of PD patients. The authors compared the SVM with three other machine learning methods, including Gaussian Naive Bayes (GNB), k-nearest neighbours (KNN), and decision tree (DT) classifiers. Since [34] did not mention the name and values of the classification hyper-parameters, we used values that are commonly optimized for these classifiers and reported them in Table 4. We implemented the models using scikit-learn v1.1.3, a reference Python library for machine learning, using Python v3.10.4.

3.3.1 Model Selection and Evaluation using Bootstrap

In this section, we describe the model selection and evaluation technique for the reproducibility experiment. [34] used a bootstrap approach to compare classifiers and optimize their hyper-parameters. To reproduce this process, we split the dataset into training (100 patients) and test (44 patients) sets having matched the HYS scores of the patients in each set. We implemented model selection using 100 iterations of a bootstrap sampling loop applied to the training set. Each iteration randomly selects (with replacement) 50 patients, normalizes the features for these 50 patients, fits the models to these 50 patients, and measures the AUC of the models on the remaining patients. To measure the stability of the models across the 100 bootstrap samples, we computed the RSD defined as:

$$RSD = \frac{\sigma_{AUC}}{\mu_{AUC}} \times 100$$

where σ_{AUC} and μ_{AUC} are the standard deviation and mean of the AUC values obtained on the 100 bootstrap samples. Finally, we selected the model with the lowest RSD and applied it to the test set.

Model	**Hyper-parameter**	**Range**
SVM	Regularization parameter	0.1, 1, 10, 100, 1000
	Gamma	1, 0.1, 0.01, 0.001, 0.0001
	Kernel type	Linear, Poly, RBF
Decision Tree	Max depth of tree	1, 2, 3, 4, 5, 8, 16, 32
	Max number of leaf nodes	2, 3, 4 , . . . , 19
	Min samples to split node	2, 3, 4, 5, 8, 12, 16, 20
K-nearest neighbors	Number of neighbors	1, 2, 3, . . . , 30
	Power parameter	1, 2
	Weight function	uniform, distance
Gaussian NB	Distribution variance	`np.logspace(0,-9, num=100)`

Table 4: Hyper-parameter grid

3.3.2 Model Selection and Evaluation using Cross-Validation

In this section, we describe the model selection and evaluation technique for the replicability experiment. We implemented a Stratified K-fold cross-validation (CV) loop similar to the one in [9] and more common than the bootstrap loop mentioned previously. We first split the cohort into training (100 patients) and test (44 patients) sets randomly. For model selection, we applied to the training set a CV loop including 50 repetitions of a 5-fold stratified CV. For each fold, we normalized the features using the standard scaler mentionned above, selected hyperparameters based on the performance of the validation set and reported the AUC computed on the test set using the model that performed the best average AUC in the validation fold. We implemented this CV loop independently for the SVM, GNB, kNN, and DT, with a scikit-learn validation pipeline, using the RepeatedStratifiedKFold function with 5 splits and 50 repetitions, and the GridSearchCV function with the parameters in Table 4.

3.4 Infrastructure & code availability

We used Pandas v.1.4.3 and Numpy v1.22.4 to construct the cohorts. The extraction of WM using SPM12 was carried out using Docker containers and Boutiques v0.5.25. The FreeSurfer volumes were extracted using a Slurm script running on a Compute Canada cluster. All the work was conducted using Ubuntu OS 22.04.

All our methods are available in a publicly available notebook (https://github.com/LivingPark-MRI/shu-etal). To comply with PPMI's Data Usage Agreements that prevent users to re-publish data, the notebook queries and downloads data directly from PPMI. Since PPMI does not have a data access API, we developed our own Python interface to PPMI using Selenium, a widely-supported Python library to automate web browser navigation. Using this interface, the notebook downloads PPMI study and imaging files to build the cohorts and train the ML models. The utility functions to download and manipulate PPMI data are merged in LivingPark utils, a Python package available on GitHub (https://github.com/LivingPark-MRI/livingpark-utils).

Chapter 4

Results

In the previous chapter, we discussed the various methods used for conducting the reproducibility and replicability experiments. Now, in this chapter, we will shift our focus towards the results that were obtained in the cohort construction, feature extraction as well as the outcomes of both experiments.

4.1 Cohorts

Table 5 summarizes the demographics of the reproduction and replication cohorts that we built. Although we built the VRC using the same PPMI filters as in [34], we were not able to reproduce the original cohort due to a shortage of subjects scanned with a Verio scanner. In fact, when we performed the query, we found a total of 29 visit pairs for progressive patients with HYS=1, 98 visit pairs for progressive patients with HYS=2, 0 visit pairs for stable patients with HYS=1 and 66 visit pairs for progressive patients with HYS=2. Using these visit pairs, we were not able to match the number of patients in [34] while ensuring that a given patient appears in at most one group.

The SRC is the closest cohort we were able to build to [34]‘s. As in [34], the SRC includes 72 progressive and 72 stable patients scanned with a Siemens manufactured MRI machine. However, the patient breakdown by HYS value differs from [34] in each group: in the SRC, both groups have 32 patients with baseline HYS=1 and 40 patients with baseline HYS=2 whereas in [34] these numbers are respectively 47 and 25. The age and F/M balance in the SRC are comparable to [34]'s with 29 females and 43 males per group, an average age of the stable group of 61.0±8.8, and an average age of the progressive group of 61.1±8.6. Finally, it should be noted that out of 144 patients in the SRC, 96 have a different value of PDSTATE

(ON/OFF) at their baseline and follow-up visits.

The MRSC includes all the patients meeting the SRC's inclusion criteria except for the MRI machine used. In the MRSC, 132 patients have been scanned with a Siemens machine, 4 with a GE Medical Systems machine, 8 with a Philips machine and 1 with an unknown machine. There are 29 females and 43 males per group. The average age of the stable group is 60.7 ± 9.4 and the average age of the progressive group is 60.7 ± 9.3. Both groups have 40 patients with baseline HYS=1 and 32 patients with baseline HYS=2. Finally, 103 patients have a different value of PDSTATE (ON/OFF) at their baseline and follow-up visits.

The NFRC included patients with an MRI of any field strength and slice thickness between 1 mm and 1.2 mm. In the NFRC, 108 patients have been scanned with a Siemens machine, 19 with a GE Medical Systems machine, 15 with a Philips machine and 2 with unknown scanners. There are 35 females and 37 males per group. The average age of the stable group is 62.0 ± 9.4 and the average age of the progressive group is 62.0 ± 9.2. Both groups have 40 patients with baseline HYS=1 and 32 patients with baseline HYS=2. Finally, 100 patients out of the 144 have a different value of PDSTATE (ON/OFF) at their baseline and follow-up visits.

The FSC includes an additional filter to only keep visit pairs with consistent values of PDSTATE (ON/OFF) between the baseline and follow-up visits. The FSC only includes 102 patients and therefore does not reproduce the sample size in the [34] cohort. In total, we could only find 22 patients with HYS=1 and 29 patients with HYS=2 in each group, totalling 102 patients in the cohort.

	Shu et al.		**VRC**		**SRC**		**MSRC**		**NFRC**		**FSC**	
	Stable	Progr	Stable	Progr	Stable	Progr	Stable	Progr	Stable	Progr	Stable	Progr
Subjects, No.	72	72	12	12	72	72	72	72	72	72	51	51
F/M No.	29/43	22/50	3/9	3/9	29/43	29/43	29/43	29/43	35/37	35/37	19/32	20/31
Age, mean +/- SD	61.30± 10.09	61.45± 11.44	66.5 ± 10.5	68.1 ± 6.9	61.0 ± 8.8	61.2 ± 8.6	60.7 ± 6.4	60.7 ± 9.3	62.0 ± 9.4	62.0 ± 9.2	60.3 ± 8.6	63.8 ± 8.9
Hoehn & Yahr Stage 1 (n)	47	47	0	0	32	32	32	32	32	32	22	22
Hoehn & Yahr Stage 2 (n)	25	25	12	12	40	40	40	40	40	40	29	29

Table 5: Summary of reproduction and replication cohorts constructed.

4.2 Feature Extraction

The second set, **RF2**, consists of the top 7 features selected by the mRMR algorithm applied to the 56 features available in PyRadiomics. We used the mRMRe v2.1.2 R package [11] with R v4.2.1.

Cohort	**RF2**
SRC	original_glrlm_LongRunLowGrayLevelEmphasis
	original_glcm_Idn
	original_glcm_ClusterShade
	original_glrlm_GrayLevelNonUniformity
	original_glszm_SizeZoneNonUniformityNormalized
	original_glcm_ClusterProminence
	original_glcm_Imc2
MSRC	original_glcm_InverseVariance
	original_glcm_JointEnergy
	original_glcm_MCC
	original_glszm_LargeAreaHighGrayLevelEmphasis
	original_glszm_SizeZoneNonUniformityNormalized
	original_glcm_ClusterShade
	original_glszm_SmallAreaLowGrayLevelEmphasis
NFRC	original_glszm_LowGrayLevelZoneEmphasis
	original_glszm_LargeAreaHighGrayLevelEmphasis
	original_glszm_SizeZoneNonUniformityNormalized
	original_glcm_ClusterShade
	original_glcm_Imc1
	original_glcm_InverseVariance
	original_glcm_Autocorrelation

Table 6: Feature extraction (**RF2**) per cohort using mRMRe (K=7)

The cluster shade (original_glcm_ClusterShade) and Size-Zone Non-Uniformity Normalized (original_glszm_SizeZoneNonUniformityNormalized) appear in all three cohorts. Notably, none of the features extracted from [34] appear in any of the cohorts.

4.3 Machine Learning Model

4.3.1 Reproducibility experiment

Our first objective was to reproduce [34]'s pipeline. For this reproducibility experiment, we used the SRC since it is the closest cohort to [34] that we could create. We used the 5 radiomic features (RF1) extracted with PyRadiomics. The distribution of RF1 in the SRC is illustrated in Figure 4. The ShortRunLowGrayLevelEmphasis feature for stable patients stood out due to the presence of several outliers, particularly on the left-tail. Unfortunately, we do not have access to [34]'s feature values, so we cannot directly compare the distribution of feature our values to theirs. The box plot displays a left-skewed distribution which suggests that the data may not be normally distributed. We trained the 4 models mentioned previously (SVM, kNN, GNB,DT), optimized the hyperparameters as described in Table 4, and used the bootstrap evaluation approach.

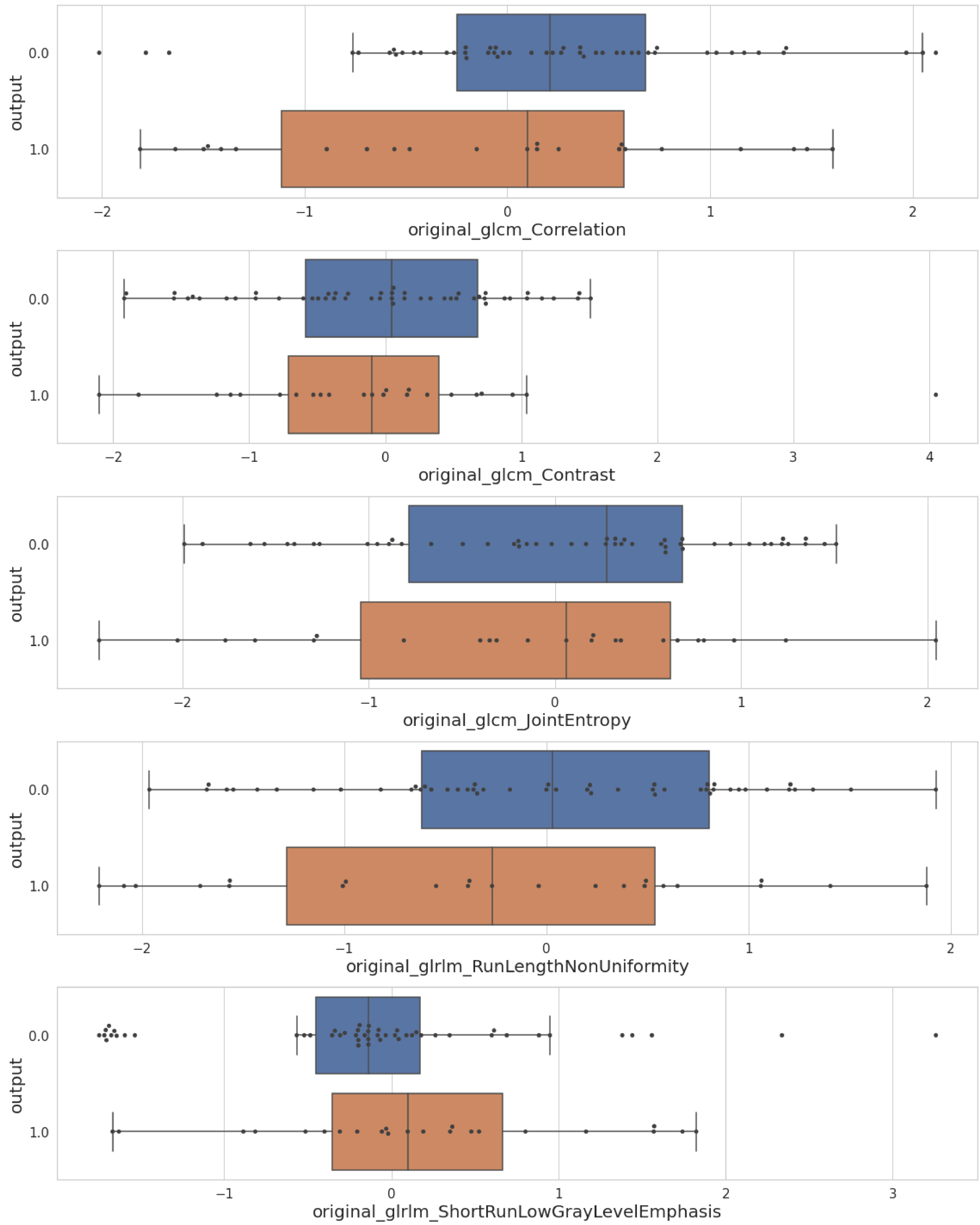

Figure 4: Box plot of the distribution of the RF1 in the training set in the SRC, with the x-axis representing the normalized feature and the y-axis representing the group (0 - stable, 1 - progressive). The features exhibit a left-skewed distribution overall, with the ShortRunLowGrayLevelEmphasis feature showing several outliers in the left-tail for the stable group.

The evaluation results for all the models are reported in Table 7. After hyper-parameter tuning, none of the models achieved an average AUC higher than 0.501 on the validation set. The Decision Tree model had the lowest RSD and was therefore selected for evaluation on the test set (with hyperparameters max depth=1, max leaf nodes=2, min samples split=2) on which it achieved an AUC of 0.523, which is slightly better than the AUC of a random classifier. Using this pipeline described we were not able to reproduce the AUC of 0.795 reported in [34].

Figure 5 shows the ROC of every model in the validation set. The observed average AUCs reach a maximal value of 0.501, suggesting that the [34] results may be the result of a random sampling artifact.

	AUC	**RSD**
SVM	0.456	10.819
Decision Tree	0.473	**7.443**
kNN	0.501	10.182
Gaussian NB	0.441	12.387

Table 7: AUC & RSD values of each model's best performer (defined as model with lowest RSD) in the validation set. The Decision Tree achieved the lowest RSD and was therefore selected for the test set in which it achieved an AUC of 0.523.

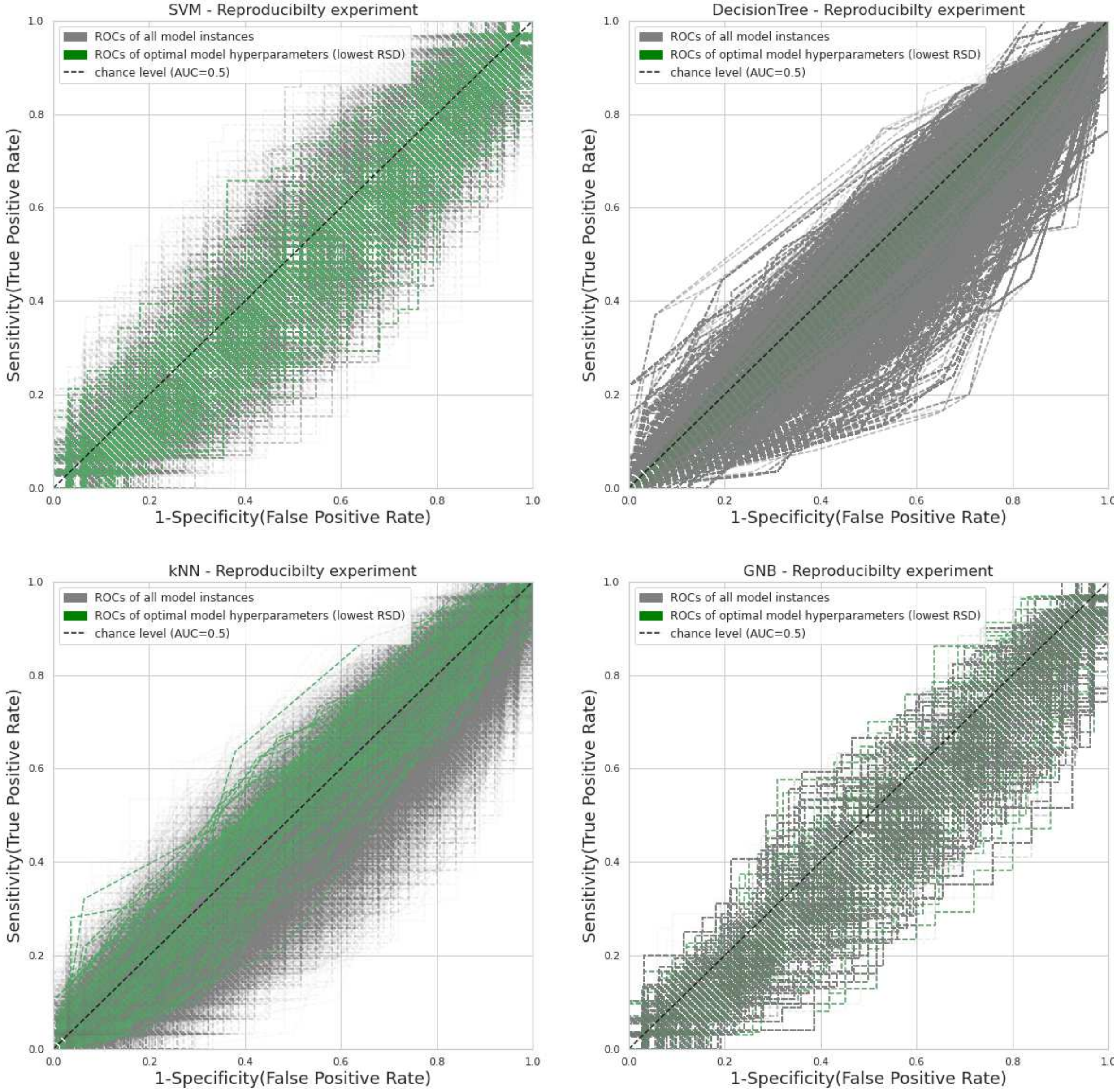

Figure 5: ROC curves per classifier. Gray lines represent ROC curves for model instances over different bootstrap iterations and hyper-parameters. The green curves show the 100 iterations of a hyper-parameter configuration that produced the lowest RSD per classifier.

4.3.2 Replicability experiment

Our second objective was to test the replicability of [34] by using several cohorts and feature sets. For every cohort we built, we trained our four models with our feature sets (RF1, RF2 and ROI volumes) using the repeated stratified K-fold CV loop. In total, we trained 3 feature sets per model resulting in 12 sets of results per cohort that were compared with [34]'s reproduced pipeline. The evaluation results for all the models are reported in Table 8.

Figure 6 shows the ROC of these configurations.

	SRC			**MSRC**			**NFRC**		
	RF1	**RF2**	**ROI**	**RF1**	**RF2**	**ROI**	**RF1**	**RF2**	**ROI**
SVM	0.54	0.59	0.6	0.629	0.354	0.513	0.42	0.634	0.374
DT	0.418	0.433	0.533	0.486	0.524	0.379	0.521	**0.685**	0.561
kNN	0.387	0.536	0.498	0.536	0.579	0.453	0.652	0.626	0.652
GNB	0.472	0.437	0.434	0.609	0.557	0.248	0.484	0.416	0.444

Table 8: Performance of the replicability experiment on the test set with AUCs according to the cohort, model and feature. Observations show that a few configurations (highlighted in green) have promising results. However, no configuration came close to [34]'s result of AUC=0.795. For the most part, the results are under chance level (AUC=0.5). There is high variability of AUC values across cohorts. RF1 and RF2 are fairly consistent.

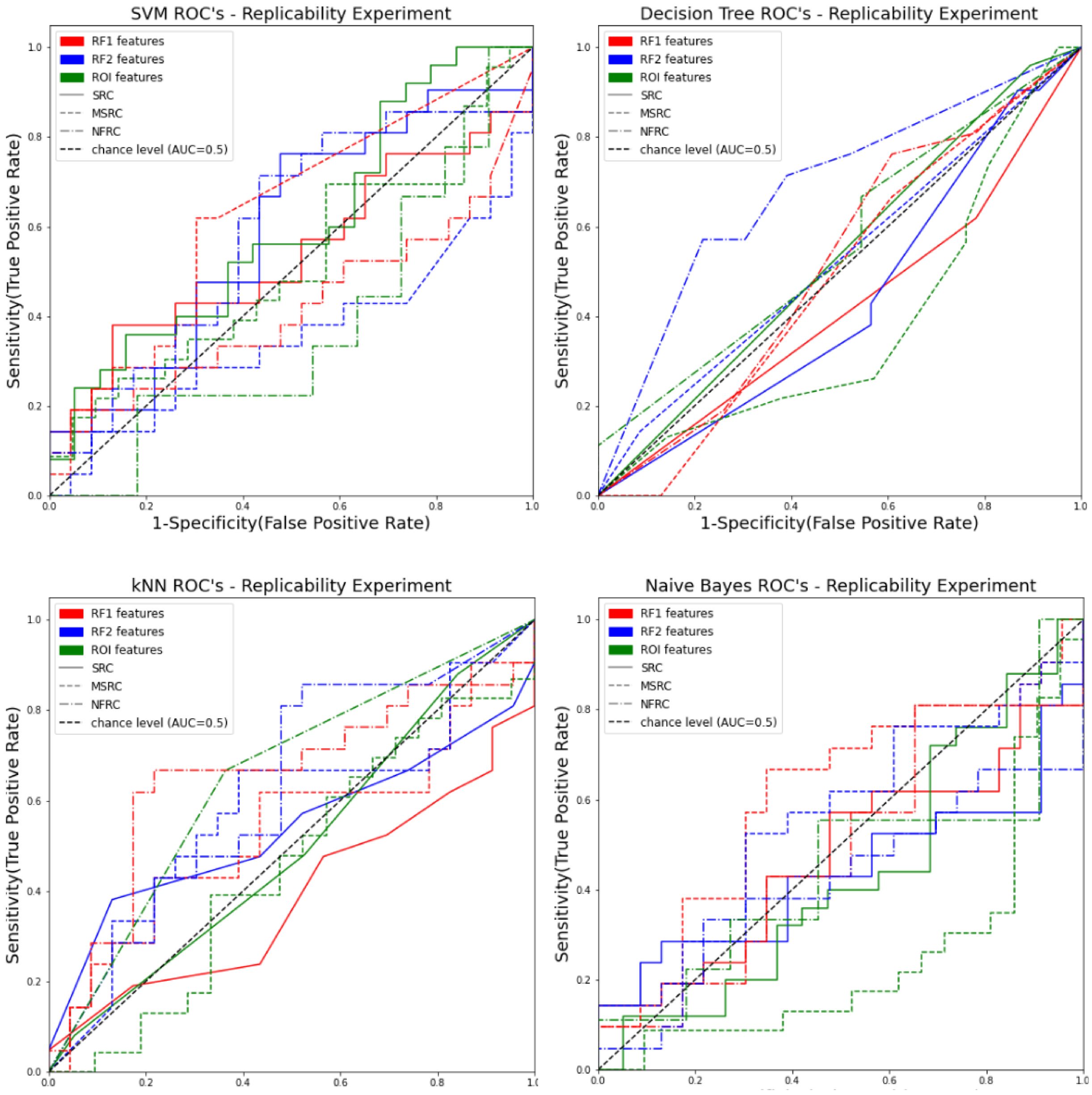

Figure 6: ROC's of replicability experiment across all cohorts and feature sets on the test set. Red = RF1 , Blue = RF2 , Green = ROI. ; Straight line = SRC, Dotted line = MSRC, Dashdot = NFRC

Chapter 5

Discussion

In the previous chapter, we presented the outcomes of our reproducibility and replicability experiments. In this chapter, we will discuss the results that we obtained in each step of the pipeline. Additionally, we will summarize our findings and provide a conclusion based on the results of all of our experiments.

In this study, we investigated the reproducibility and replicability of a recent study applying ML to MRI data to predict PD progression. We attempted to reproduce the findings in [34], but were unable to do so despite following the same methods and using data from the same database. The failure to reproduce these results were related to the three phases of the analysis: cohort construction, feature extraction, and ML model construction.

The cohort construction proved to be more challenging than expected. In Section 3.1, we attempted to reproduce the exact cohort used in [34] by applying the same filters as described in their methods. Instead, the SRC was used as the main reference cohort since it closely resembled [34]'s cohort. The major difference between [34]'s cohort and the SRC was the baseline HYS score distribution per group. Our split for baseline HYS scores of 1 and 2 per group is 32/40 respectively, whereas [34] reports a split of 47/25 respectively. There is a considerable difference between the SRC's distribution and [34]'s cohort that we could not ignore. The disparity in HYS scores between both cohorts has the potential to influence the performance of our ML models. A possible reason why we could not reproduce the cohort is that there may have been participants that withdrew from PPMI's study. We initially thought that this would result in their data disappearing from PPMI. However, according to PPMI's protocols section 22, any data collected before a subject's withdrawal will not be removed. Another possible explanation for our failure to reproduce [34]'s cohort could be due to changes in the PPMI user interface that might be due to clerical errors. PPMI's advanced image search tool has gradually enabled users to input increasing numbers of filters over time. We believe that the introduction of the "Mfg Model" (manufacturer model) filter impacted our ability to search the database, as not all patient metadata may have been updated with the Mfg Model information. There may be a significant number of patients scanned with a Siemens manufacturer MRI machine without Mfg Model information. After investigating PPMI's database, we found that several images labeled as Proton Density weighting (PDw) are actually T1-weighted images. Further analysis revealed that there are approximately 800 mislabeled T1-weighted MRIs labeled as PDw images in the imaging database. More details on this fix can be found in the LivingPark utils v0.9. For the MSRC and NFRC, there were no difficulties in constructing these cohorts as the increasingly permissive filters allowed for the inclusion of more patients. Finally, despite our efforts to impute missing data in the MDS-UPDRS Part III evaluation, the FSC did not meet the requirements set by [34]. In summary, the cohort construction was difficult and our attempt to reproduce [34]'s cohort mostly failed.

We reproduced the pre-processing pipeline after making necessary modifications. The WM extraction using SPM12 was straightforward. Once we collected the WM, we did not follow the manual steps outlined by [34]. Instead, we performed QC on the WM by manually inspecting every image using 3D Slicer v5.0.3, while [34] used ITK-snap for QC. We do not believe that the difference in software (ITK-snap and 3D Slicer) is a major source of variability. However, the variation in the QC protocol between our study and [34] could potentially introduce variability in the radiomic features. The most challenging aspect was reproducing the feature extraction step as the A.K software is not publicly accessible and we could not obtain access to it. We used PyRadiomics instead to extract radiomic features. [34] extracted 378 features while we extracted 56 features. The large difference in the number of features is due to the fact that PyRadiomics calculates the value of a feature for each angle separately, after which the mean of these values is returned. Not having the exact same features as [34] could have a big impact on the ML models. Nevertheless, we were able to extract radiomic features which was a critical step of the pipeline. For feature selection, we used the mRMRe library to select the top 7 ranked features, which were different from those used in [34]. The ROI volumetric data was easy to extract thanks to the availability of FreeSurfer. Although the ROI features extracted in (Chougar et al., 2020) are not the primary focus of our study, we found that the ROI volumes enhance the replicability aspect of this study.

The model we attempted to reproduce in [34] yielded underperforming results. Using the

SRC, RF1 and a bootstrap evaluation method, we achieved AUC's that are below chance level (0.5). The Decision Tree model proved to be the most stable with an RSD of 7.443 with maximum depth = 1, maximum number of leaf nodes = 2 and minnimum sample split = 2 as hyperparameters. The kNN model achieved an AUC of 0.501 in the validation set, which was the highest among all models tested. The results suggest that the models are unable to distinguish between stable and progressive patients. It seems to be making random or constant predictions for all patients. We believe this is due to the very limited number of training samples (n=50) in the bootstrap method which results in an insufficient amount of data for the model to differentiate each group. Our reproducibility experiment may have implications for ML research in the context of PD as well as for how results are reported. The use of unconventional evaluation methods and the inevitable lack of methodological details in ML/MRI papers is likely an important factor explaining our failure to reproduce the results. It is important for researchers in the field of ML and PD to ensure that their methods are well documented, available publicly, and make use of standard techniques and tools. We tried to communicate with the authors of [34] by sending them two emails to discuss our results and seek clarification on why we couldn't reproduce their study. Unfortunately, we have not received any reply from them yet.

To improve reproducibility in ML research for PD progression prediction, we make suggestions for the community to consider when publishing their work. First and foremost, publishing re-executable notebooks is highly recommended to allow fellow researchers to follow-up on one's work. When it comes to feature extraction, one should use publicly-available tools such as FreeSurfer, SPM12, or PyRadiomics instead of proprietary softwares. For ML related techniques, having a predefined training and test set can introduce bias and should be avoided. Instead, it is preferable to randomly split the training and test set, and to never peek at the test set until training is complete. For model evaluation, while bootstrap is not necessarily bad practice, it is definitely not as widely recognized as CV. We strongly recommend the proper usage of CV with well-defined hyperparameter tuning to thoroughly evaluate a model's performance. The work in [30] proposed an ML checklist introduced at the NeuriIPS 2019 reproducibility program that can be used for ensuring good practices of ML techniques. Moreover, [27] provides a user-friendly interactive checklist for neuroimaging guidelines (https://ohbm.github.io/eCOBIDAS/#/). The COBIDAS protocol offers users a list of considerations that researchers can refer to when conducting a neuroimaging analysis pipeline, such as MRI acquisition, preprocessing, etc.

The best performer in the replicability experiment was obtained with the NFRC cohort trained with a decision tree (using hyper-parameters max depth = 8, max leaf nodes = 19 and min sample split = 3) with RF2 features, achieving an AUC of 0.685 on the test set. Although some models achieved performances above chance level, none of the models came close to the results reported by [34]. The replicability experiment results showed great variability between cohorts, feature sets and models.

The results across cohorts lacked consistency. For instance, the kNN with RF1 achieved AUC values of 0.387, 0.536, and 0.652 across the SRC, MSRC and NFRC respectively. Large differences between cohorts were found. For instance, the AUC of SVM using ROI features was 0.374 with NFRC and .6 with SRC. The SRC, MSRC and NFRC averaged AUC values across models and features of 0.488, 0.489 and 0.539 respectively. We were surprised to find that the NFRC performed better than the SRC, given that the SRC has patients scanned from the same MRI manufactured machine. We expected the SRC to have an advantage since the radiomic features extracted from the same MRI manufactured machine are more likely to be consistent.

For some iterations in the CV loops, one radiomic feature set performed better than the other, however, the results were overall consistent. The average AUC values of RF1 and RF2 were 0.513 and 0.531 respectively. The features selected in RF2 (see Table 6) across cohorts were very different from those selected by [34] in RF1, however, the marginal difference in average AUC values suggest that there is negligible variability in radiomic feature selection. The average AUC value of ROI features was 0.472, however, some configurations revealed promising results (0.6 with SRC & SVM, and 0.652 with NFRC & kNN). These results suggest that ROI features could potentially be useful biomarkers of PD progression detection, further validation is necessary to investigate its utility.

This work has several limitations to be considered. [34] developed their ML models with R v3.5.1. In principle, we should have implemented our models using R for the reproducibility experiment. We chose to use Python instead of R since in our experience the two languages yield similar computational results. The imputation process used for missing scores in the MDS-UPDRS Part III evaluation may contain errors due to the manual derivation of rules. Consequently, this could result in the FSC being unsuccessful as there may not be enough patients to construct the cohort. Moreover, negative bias in replication studies refers to the pattern of replication studies that report weaker results compared to original studies. To address negative bias, we attempted many different replications (as seen in Figure 3) to ensure the robustness of our findings.

Our study has promising opportunities for future work. For feature selection, we could opt to use CNN's to learn features instead of using radiomic features. Alternatively, we could further validate the performance of ROI volumes, as we saw their potential utility in the replicability experiment. Further validation of this feature set would be an interesting focus for future research. Since our models could not properly predict disease progression, we could increase cohort sizes to improve model training. For feature extraction, we could explore additional feature selection techniques such as Principal Component Analysis or Random Forests. All in all, there are many possible replications in [34]'s pipeline that could be interesting directions for future research.

To conclude, while we were unable to replicate [34]'s study with our adapted pipeline, we found that there is a potential for brain imaging biomarkers for PD progression prediction. Future work will likely need more data to ensure robust and generalizable results across cohorts and imaging acquisitions.

Appendix A

Literature review of MRI-derived biomarkers for PD

Title	Objectives	Target variable	Software	ML models
Complex networks reveal early MRI markers of Parkinson's disease.	Propose an unsupervised general methodology to model brain connectivity for PD & NC. Explore regions affected by disease (ROI). Propose a learning strategy to combine CF & NF. Create diagnostic tool for PD diagnosis based on MRI features. The proposed approach both detects which regions are mostly affected by the disease and uses the network measures to provide classification scores.	Diagnosis (PD vs HC)	FSL, Brain Extraction Tool, FLIRT, FreeSurfer, R 3.2.2	Random Forests

Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease.	Identify ROIs using VBM & analyse regions for PD detection. Extract features based on first and second order statistics. Select features based on PCA.	Diagnosis (PD vs HC)	CAT12, SPM12, JULIA 1.4, WEKA	kNN (k=5), MLP, SVM radial kernel, Random forest, NB, LC, Bayesian network
An Integrative Nomogram for Identifying Early-Stage Parkinson's Disease Using Non-motor Symptoms and White Matter-Based Radiomics Biomarkers From Whole-Brain MRI.	Segment WM to extract radiomic features and develop radiomic biomarkers. Combine biomarkers with non-motor symptoms to build a nomogram. The proposed approach suggests the possibility of developing novel imaging biomarkers of PD from WM using radiomics and combining them with prodromal nonmotor symptoms to generate an integrative nomogram for disease classification.	Diagnosis (PD vs HC)	SPM12, ITK-SNAP, AK software, SPSS 22.0 , GraphPad Prism 6, R 3.3.1	SVM, Bayes, Logistic Regression, Random Forests, Decision Trees

Predicting the progression of Parkinson's disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter.	The proposed approach develops a radiomics method using whole-brain WM based on a machine learning approach to detect disease stage and progression to PD in patients using conventional MRI.	Stable PD vs Progressive PD	SPM12, ITK-SNAP, AK software, SPSS 25.0, R 3.5.1, Python 3.5.6	SVM (Linear kernel), Bayes, kNN, DT
Sparse feature learning for multi-class Parkinson's disease classification.	Propose a framework to construct a LSR model based on LDA and LPP. Build a multiclassifier for PD using PD, HC, SWEDD scans	Diagnosis (PD vs HC)	libsvm5, FSL	SVM
Automated Categorization of Parkinsonian Syndromes Using Magnetic Resonance Imaging in a Clinical Setting	The proposed approach is to asses the predictive performance of ML for categorisation of Parkinsonian symptoms.	PD, MSA-C, MSA-P, atypical pakinsonism	Matlab R2017b, FreeSurfer 6.0, FMRIB 5.0, SPM12, scikit-learn	Logistic regression, SVM, Random Forest

Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of Parkinson disease.	Combines KSOM and LSSVM for PD detection. Compares PD, NC, SWEDD scans. KSOM is used as a vector quantization technique to reduce complexity. LS-SVM is employed for classification of subjects & PD, HC, SWEDD.	Diagnosis (PD vs HC vs SWEDD)	VBM8, SOMtoolbox, Matlab, BrainNet Viewer	KSOM, LSSVM
Detection of Parkinson's Disease from 3T T1 Weighted MRI Scans Using 3D Convolutional Neural Network.	Develop a 3D CNN model for PD detection. 3D MRI analysis was performed for the detection of PD using 3D convolutional neural network and PD, HC.	Diagnosis (PD vs HC)	Python	CNN

Joint feature-sample selection and robust diagnosis of Parkinson's disease fromMRI data.	Create a joint feature-sample method to select optimal samples & features. Create a robust classification model to de-noise the selected subset of features and samples. propose a new joint feature-sample selection (JFSS) procedure,which jointly selects the best subset of most discriminative features and best samples to build a classification model. Robust classification framework is proposed to simultaneously de-noise the selected subset of features and samples. Diagnosis (PD vs HC)	Diagnosis (PD vs HC)	Not shared	LDA, MC, SVM, SR, JFSS-C

Determination of Imaging Biomarkers to Decipher Disease Trajectories and Differential Diagnosis of Neurodegenerative Diseases (DIsease TreND).	Detect imaging biomarkers & build automated disease diagnosis using SOM & LS-SVM.	Diagnosis (AD, HC, MCI, PD, SWEDD)	SPM 8.0, VBM8	LS-SVM
Effectiveness of imaging genetics analysis to explain degree of depression in Parkinson's disease.	Imaging genetic features predicted and explained the degree of depression in Parkinson's disease appropriately	Depressed PD vs Non-depressed PD	NeuroX, FSL	LASSO

Table 9: Preview of the literature review performed consisting a paper's title, main objectives, target variable, software used and ML models trained. To view the full grid, refer to this google document.

Appendix B

Data cleaning rules implemented in MDS UPDRS III

	Inconsistency
Problem:	Some records have missing data for all UPDRS-III variables.
Fix:	Remove these records
Problem:	A few records have PDSTATE=ON and PDTRTMNT=0, which is inconsistent:
Fix:	Set PDTRTMNT=1 for these records. It doesn't seem realistic that a plausible PDMEDTM and PDSTATE=ON have been entered by mistake while the patient was not under medication.
Problem:	Some records have a non-empty PDMEDTM and have PDTRTMNT=0, which is inconsistent.
Fix:	Set PDTRTMNT=1. It is unlikely that a plausible medication time was entered by mistake.
Problem:	Some patients were on medication at screening time while PPMI patients were supposed to be unmedicated at screening time.
Fix:	Keep the records. Maybe the patients started medication between recruitment and screening time.

Inconsistency

Problem: Some records have PDSTATE=ON but PDMEDTM is after EXAMTM.

Fix: Discard the records.

Problem: Some visits have 3 exams while a maximum of two exams per visit are expected, one in OFF state and one in ON state.

Fix: Remove exam with EXAMTM=NaN and PDSTATE=NaN when visit has 3 exams

Appendix C

Imputation of missing PDSTATE & PDTRTMNT in MDS UPDRS III

	Case
Case I	PDSTATE=OFF and PDTRTMNT=NaN.
Fix:	Set PDTRTMNT=0. It is unlikely that these records correspond to medicated patients when none of the variables related to medication have a value.
Case II	PDSTATE=NaN and PDTRTMNT=0.
Fix:	Set PDSTATE=OFF. The patient is not medicated and for this reason PDSTATE is likely to not have been entered.

Case

Case III PDSTATE=NaN and PDTRTMNT=1

Fix: **IF** record belongs to a visit with two exams:

- Set PDSTATE=OFF for record with earliest EXAMTM.
- Set PDSTATE=ON for record with latest EXAMTM.

ELSE determine PDSTATE as a function of PDMEDTM and EXAMTM:

- **IF** PDMEDTM or EXAMTM are missing **THEN** discard records.
- **IF** PDMEDTM is earlier than EXAMTM
  - **IF** PDMEDTM is earlier than EXAMTM
    - **IF** EXAMTM-PDMEDTM $\geq$ 30 min **THEN** set PDSTATE=ON.
    - **ELSE** Discard record
  - **ELSE** set PDSTATE=OFF

Bibliography

[1] Leaning into the replication crisis: Why you should consider conducting replication research.

[2] Parkinson disease.

[3] Association for Computing Machinery. Artifact Review and Badging., 2021. [Online; accessed 6-October-2021].

[4] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

[5] Lorena A. Barba. Terminologies for reproducible research. *CoRR*, abs/1802.03311, 2018.

[6] Roongroj Bhidayasiri and Daniel Tarsy. Parkinson's disease: Hoehn and yahr scale. *Current Clinical Neurology*, page 4–5, 2012.

[7] Sabyasachi Chakraborty, Satyabrata Aich, and Hee-Cheol Kim. Detection of parkinson's disease from 3t t1 weighted mri scans using 3d convolutional neural network. *Diagnostics*, 10(6), 2020.

[8] Andrea Cherubini, Maurizio Morelli, Rita Nisticó, Maria Salsone, Gennarina Arabia, Roberta Vasta, Antonio Augimeri, Maria Eugenia Caligiuri, and Aldo Quattrone. Magnetic resonance support vector machine discriminates between parkinson disease and progressive supranuclear palsy. *Movement Disorders*, 29(2):266–269, 2013.

[9] Lydia Chougar, Johann Faouzi, Nadya Pyatigorskaya, Lydia Yahia-Cherif, Rahul Gaurav, Emma Biondetti, Marie Villotte, Romain Valabrègue, Jean-Christophe Corvol, Alexis Brice, and et al. Automated categorization of parkinsonian syndromes using magnetic resonance imaging in a clinical setting. *Movement Disorders*, 36(2):460–470, 2020.

[10] Jon F. Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. *SEG Technical Program Expanded Abstracts 1992*, 1992.

[11] Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. Mrmre: An r package for parallelized mrmr ensemble feature selection. *Bioinformatics*, 29(18):2365–2368, 2013.

[12] George DeMaagd and Ashok Philip. Parkinson's disease and its management: Part 1: Disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis, Aug 2015.

[13] Daniele Fanelli. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11):2628–2631, 2018.

[14] Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012.

[15] Schloss Dagstuhl Leibniz Center for Informatics. Home, Feb 2023.

[16] Thiago FA França and José Maria Monserrat. Reproducibility crisis in science or unrealistic expectations? *EMBO reports*, 19(6), 2018.

[17] 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.

[18] Ed H. Gronenschild, Petra Habets, Heidi I. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, and Machteld Marcelis. The effects of freesurfer version, workstation type, and macintosh operating system version on anatomical volume and cortical thickness measurements. *PLoS ONE*, 7(6), 2012.

[19] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018.

[20] John P. Ioannidis, Richard Klavans, and Kevin W. Boyack. Thousands of scientists publish a paper every five days. *Nature*, 561(7722):167–169, 2018.

[21] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, Mark W. Woolrich, and Stephen M. Smith. Fsl. *NeuroImage*, 62(2):782–790, 2012.

[22] Gary King. Replication, replication. *PS: Political Science and Politics*, 28:444–452, September 1995. See updates to this paper&nbsp;for how I use this paper as a class assignment now.

[23] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, and et al. The parkinson progression marker initiative (ppmi). *Progress in Neurobiology*, 95(4):629–635, 2011.

[24] Trina Mitchell, Stéphane Lehéricy, Shannon Y. Chiu, Antonio P. Strafella, A. Jon Stoessl, and David E. Vaillancourt. Emerging neuroimaging biomarkers across disease stage in parkinson disease. *JAMA Neurology*, 78(10):1262, 2021.

[25] Marcin Miłkowski, Witold M. Hensel, and Mateusz Hohol. Replicability or reproducibility? on the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience*, 45(3):163–172, 2018.

[26] Congjun Mu. Replication research in applied linguistics. graeme porte (ed.). new york: Cambridge university press, 2012. pp. xvi 286. *Studies in Second Language Acquisition*, 35(4):763–764, 2013.

[27] Thomas E. Nichols, Samir Das, Simon B. Eickhoff, Alan C. Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, Michael P. Milham, Russell A. Poldrack, Jean-Baptiste Poline, Erika Proal, Bertrand Thirion, David C. Van Essen, Tonya White, and B. T. Thomas Yeo. Best practices in data analysis and sharing in neuroimaging using mri. *bioRxiv*, 2016.

[28] Ariane Park and Mark Stacy. Non-motor symptoms in parkinson's disease. *Journal of Neurology*, 256(S3):293–298, 2009.

[29] Laia Pedro-Roig and Christoph H Emmerich. The reproducibility crisis in preclinical research – lessons to learn from clinical research. *Medical Writing*, 26:28–32, 2017.

[30] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (A report from the neurips 2019 reproducibility program). *CoRR*, abs/2003.12206, 2020.

[31] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712, 2011.

[32] Mohammad R. Salmanpour, Mojtaba Shamsaei, Ghasem Hajianfar, Hamid Soltanian-Zadeh, and Arman Rahmim. Longitudinal clustering analysis and prediction of parkinson's disease progression using radiomics and hybrid machine learning. *Quantitative Imaging in Medicine and Surgery*, 12(2), 2021.

[33] C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M.C. Gilardi, A. Quattrone, and et al. Machine learning on brain mri data for differential diagnosis of parkinson's disease and progressive supranuclear palsy. *Journal of Neuroscience Methods*, 222:230–237, 2014.

[34] Zhen-Yu Shu, Si-Jia Cui, Xiao Wu, Yuyun Xu, Peiyu Huang, Pei-Pei Pang, and Minming Zhang. Predicting the progression of parkinson's disease using conventional mri and machine learning: An application of radiomic biomarkers in whole-brain white matter. *Magnetic Resonance in Medicine*, 85(3):1611–1624, 2020.

[35] Gurpreet Singh and Lakshminarayanan Samavedham. Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of parkinson disease. *Journal of Neuroscience Methods*, 256:30–40, 2015.

[36] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, Hugo J.W.L. Aerts, and et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21), 2017.

[37] Jing Yang, Roxana G. Burciu, and David E. Vaillancourt. Longitudinal progression markers of parkinson's disease: Current view on structural imaging. *Current Neurology and Neuroscience Reports*, 18(12), 2018.

[38] Adrian Zbiciak and Tymon Markiewicz. A new extraordinary means of appeal in the polish criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment. *Access to Justice in Eastern Europe*, 6(2):1–18, March 2023.