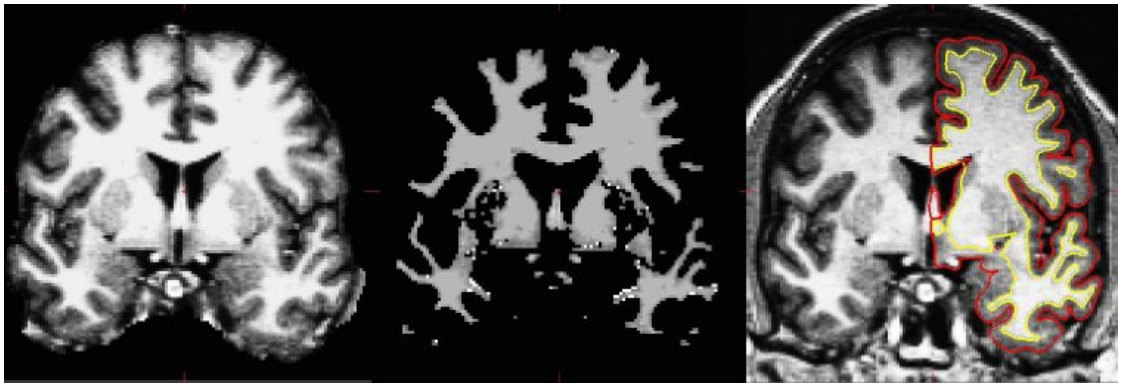
TÉCNICO LISBOA



Detection of Dopamine Deficiency for Parkinson's Disease Diagnosis with Machine Learning and Structural MRI

Carolina Isabel Matias Vicente

Thesis to obtain the Master of Science Degree in

Data Science and Engineering

Supervisor(s): Doctor Diana Prata
Doctor David Manuel Martins de Matos

Examination Committee

Chairperson: Doctor Maria do Rosário De Oliveira Silva
Supervisor: Doctor Diana Prata
Member of the Committee: Doctor Maria da Conceição Esperança Amado

October 2021

Dedicated to the ones who will read this.

Acknowledgments

I want to thank my supervisors, David Matos and Diana Prata, and all the people with whom I worked with from Diana Prata's Lab, Vasco and Helena specially, for all the help and support.

I want to thank my parents for always supporting me and my brother for annoying me but still helping me.

Last but not least, I want to thank my friends that heard me through all my university years, Mariana Carrasco for always being there and for all the late night projects, João Mira for the support and our meetings, Henrique Santos for all the company and support and Ted for the distraction and always being available when I ask.

Resumo

Doença de Parkinson tem origem neurológica e afeta 1% da população com mais de 60 anos. Múltiplas doenças têm sintomatologia semelhante, contudo a doença de Parkinson caracteriza-se pela perda de neurónios dopaminogénicos. Esta perda leva a deficiência de dopamina, que pode ser detetada usando um *DaTscan*. Sujeitos inicialmente diagnosticados com Parkinson apresentando um exame negativo são categorizados como *Scans Without Evidence of Dopamine Deficiency*. Esta dissertação estuda a possibilidade de distinguir sujeitos com e sem deficiência de dopamina utilizando Imagem por Ressonância Magnética Estrutural.

Imagens de 311 sujeitos do estudo PPMI foram processadas usando *FreeSurfer* que calculou 689 características. Os dados foram particionados em 70% e 30% para treino e teste, respetivamente. Da amostra de treino foram escolhidas observações (10%) para validação dos modelos. Um *pipeline* de Aprendizagem Automática já existente foi usado como *baseline*. Vários algoritmos foram comparados com este. Para seleção de características, as características foram separadas em conjuntos de acordo com a região do cérebro, ou alternativamente, usando as características da versão robusta da Análise de Componentes Principais.

O algoritmo *baseline* sofreu sobreajuste, apresentando uma exatidão balanceada de 96.6% e 54.5% para os dados de treino e validação, respetivamente. Todas as outras abordagens usadas resultaram em sub-ajuste ou sobreajuste, sendo a maior exatidão balanceada para os dados de validação de 80.42% e, para validação cruzada, 62.67%. No conjunto de teste, a maior exatidão balanceada foi de 50.60%.

Keywords: Doença de Parkinson, SWEDD, Aprendizagem Automática, Imagem por Ressonância Magnética Estrutural

Abstract

Parkinson's disease is a neurological disorder that affects 1% of the population over 60. Multiple diseases cause similar symptoms, but Parkinson's disease is characterized by dopaminergic neuronal loss. This leads to dopamine deficiency, which can be detected with a DaTscan. Subjects initially diagnosed with Parkinson's, but who have a negative exam are grouped as patients with Scans Without Evidence of Dopamine Deficiency. The present work aims to achieve the distinction of subjects with and without dopamine deficiency with a structural Magnetic Resonance Imaging scan.

Images from 311 subjects from the PPMI database were processed with *FreeSurfer* into 689 features. Data was then divided into 2 categories, with 70% allocated to training sets, and 30% set aside for test sets. Cross-validation and a validation set, made up of 10% of the training data, were used to compare different modelling approaches. An existing Machine Learning pipeline was used as a baseline approach. Multiple algorithms were compared. For feature selection, the features were partitioned into sets according to brain region, and as an alternative, features from robust Principal Component Analysis.

The baseline approach overfitted, with accuracies of 96.6% and 54.5% for training and validation sets, respectively. All other simpler approaches resulted in underfitting or overfitting, with the highest validation balanced accuracy being 80.42% and 62.67% for cross validation. These were tested in the independent test set where the highest balanced accuracy was 50.60%.

Keywords: Parkinson's Disease, SWEDD, Machine Learning, Magnetic Resonance Imaging

Contents

List of Tables

List of Figures

Acronyms

DaTscan DaTscan Single Photon Emission Computed Tomography. xv, 1, 6, 17, 18, 20, 24–29, 39, 40

ML Machine Learning. 1–3, 8, 9, 12, 14–16, 21, 22, 25, 37, 39

MRI Magnetic Resonance Imagining. xv, 2, 5–8, 10, 12, 13, 18, 20, 22, 24, 37, 39, 40

PCA Principal Component Analysis. xv, 12, 13, 27–29

PD Parkinson's Disease. xiii, xv, 1–6, 8–12, 15, 17, 20, 40

PPMI Parkinson's Progression Markers Initiative. 17

RPCA Robust Principal Component Analysis. xv, 13, 21–23, 28–30, 36, 39

sMRI structural Magnetic Resonance Imaging. 1, 6, 19, 36, 39, 40

sMRI T1w structural Magnetic Resonance Imaging, T1 weighted. xiii, 6, 7, 12, 17–19, 39

SWEDD Scans Without Evidence of Dopamine Deficiency. xiii, 1, 6, 8, 12, 25, 40

Chapter 1

Introduction

Parkinson's Disease (PD) is a neurological disorder that currently affects 1% of the population over 60 years old [1], and it is expected to affect a greater percentage of the population in the decades to come [2].

Multiple diseases can cause similar symptoms to PD, such as tremors, but PD in particular is characterized by a loss of dopaminergic neurons. This leads to dopamine deficiency, which can be detected using a DaTscan Single Photon Emission Computed Tomography (DaTscan). Subjects initially misdiagnosed with PD, but having a negative DaTscan are grouped as Scans Without Evidence of Dopamine Deficiency (SWEDD). Although important as a diagnostic tool, DaTscan is invasive, expensive, and not readily available.

1.1 Hypothesis and Objectives

Th hypothesis of the present work is to aim to achieve the distinction of subjects with and without dopamine deficiency with a structural Magnetic Resonance Imaging (sMRI) scan. This is something that is not found in the literature, but which would offer clinicians a tool to assist in differentiating between diagnoses. The detection and accuracy of diagnosis is important to improve patient quality of life and treatment options, while also contributing to ongoing research into treatments and cures for these diseases.

This work was done in collaboration with Diana Prata's Lab and NeuroPsyAI, who proposed this idea and provided knowledge from their previous work and research.

The main objective is to study if the hypothesis is possible. Starting by trying an already existing approach as a baseline, then exploring different algorithms, data transformations and Machine Learning (ML) methods. Validate these methods to choose the best one, to then test it in an independent set. If the hypothesis is achieved, then get some insights on how the distinction is done, for instance, what brain regions are the most important.

1.2 Thesis Outline

In chapter 2 Background, there are details of PD and of its diagnosis, along with the state of the art of the efforts to create diagnostic tools for PD with ML. Furthermore, there is an introduction to ML algorithms that are used in later chapters.

In chapter 3 Methods, the data source is explained and how the MRI are chosen, and the extracted features. Then, how the data was separated for validation and testing. Moreover, the data transformations and balancing types that are used are explained. Finally, the baseline approach is detailed and the approach used for exploring different options, along with some final considerations and ideas that were to be used if the hypothesis was proven right.

In chapter 4 Results, the results of data analysis and all the methods used are shown along with some discussion that goes into why things might not have worked, since the hypothesis was not able to be proven right.

In the final chapter 5 Conclusions, there is an overview of everything that was done, with the conclusions that can be taken from the work and suggestions for future work to try to get better results.

Chapter 2

Background

In this chapter, some background on Parkinson's Disease (PD) is discussed, an overview of Magnetic Resonance Imaging is provided, and a typical pipeline for applying ML on these images is presented. This chapter also provides an overview of relevant literature, with more detail on papers that studied similar hypothesis to the one proposed by this thesis. Finally, Principal Component Analysis and ML algorithms are introduced.

2.1 Parkinson's disease

Parkinson's Disease is a neurological disorder that affects the older population and it is expected to affect a greater percentage of the population in the decades to come [2]. It currently affects 1% of the population over 60 [1], with more than 6 million individuals affected worldwide [3]. PD is uncommon in individuals younger than 50 years old, and increases in prevalence with age, but more in men than in women [3].

In regards to the pathogenesis, most cases of PD are idiopathic (unknown cause), but there are known genetic and environmental contributing factors [3]. The origin of PD was thought to be only due to environmental risk factors, like rural living, pesticide use, and well-water consumption [4]. However, research from the last decade has pointed otherwise: currently the greatest risk factor for PD is age, and there are some environmental contributions like pesticide exposure, smoking, and caffeine intake, but none have been confirmed as causes. On the other hand, several genetic contributions to this disease have been confirmed by research [2].

Pathologically, PD is characterised by a loss of dopaminergic neurons and the presence of Lewy bodies, which are abnormal aggregations of protein that develop inside nerve cells in the midbrain [2]. This leads to non-motor symptoms, like sleep disorders, and motor symptoms, for instance tremors and rigidity, which are the most well known PD symptoms.

2.1.1 Pathophysiology

The Braak hypothesis is the most cited model that explains the PD neuropathological progression, [5]. It suggests that PD has 4 stages:

Stage 1 and 2 – PD starts in the medulla and olfactory bulb. This pathology is associated with symptoms that occur prior to the movement disorder's onset, for instance rapid eye movement (REM) sleep behavior disorder and decreased sense of smell.

Stage 3 and 4 – Pathology progresses to the substantia nigra pars compacta (SNpc), and other structures in the midbrain and basal forebrain. This is associated with the traditional PD motor symptoms, which appear once there is a loss of approximately half of the cells in the caudal substancia nigra [3, 6]. In advanced PD, the pathology progresses to the cerebral cortices, which leads to cognitive impairment and hallucinations.

Neuronal loss in PD occurs in other brain regions, including the locus ceruleus, nucleus basalis of Meynert, pedunculopontine nucleus, raphe nucleus, dorsal motor nucleus of the vagus, amygdala, and hypothalamus [6]. Moreover, Figure 2.1 shows some of the brain regions affected by PD, according to [4].

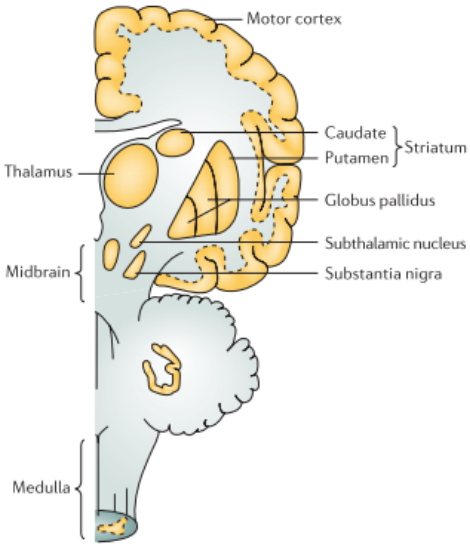


Figure 1 | **The main brain regions affected in Parkinson disease.** A lateral section through a human brain is shown, with the anterior to the left. The yellow shading indicates regions of the brain that are affected in Parkinson disease

Figure 2.1: Brain regions affected by PD [4]

Non-motor Symptons – Loss of sense smell, sleep disturbances like REM sleep behavior disorder, constipation, urinary dysfunction, orthostatic hypotension, excessive daytime sleepiness, and depression.

Motor Symptons – Tremors, stiffness, bradykinesia (slowness of movement), rigidity, and imbalance.

2.1.2 PD Diagnosis Process

There is currently no definitive test for the diagnosis of PD for a living subject. Diagnosis requires post mortem examination of the brain for neuronal loss and depigmentation of the substantia nigra, in addition to the presence of Lewy bodies in the brain stem [7, 8].

Patients look for a diagnosis when they reach a stage with motor symptoms. Prior to that patients are considered to be in a prodromal phase that may turn into PD. Parkinsonism is a general term for a group neurological disorders that cause motor symptoms such as tremors, slow movement, and stiffness. PD is the most common disease to explain a parkinsonism case, and the diagnosis process for a patient presenting symptoms of parkinsonism relies on the expertise and experience of clinicians to distinguish and identify the underlying disease. For PD diagnosis, an example of how this process is performed can be seen in Figure 2.2. The symptoms during the prodromal phase are not PD specific, but when multiple co-occur it is an indication of a subsequent PD diagnosis [3].

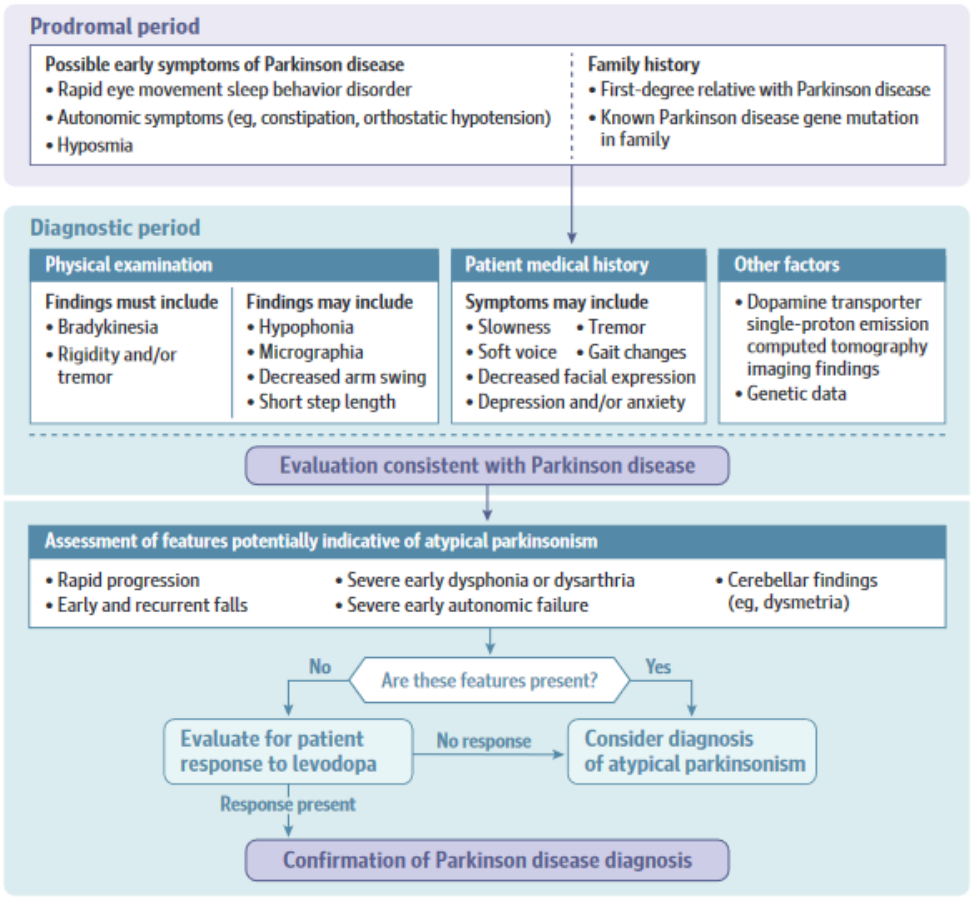


Figure 2.2: PD diagnosis process [3]

The use of medication and exams can help in the differentiation between different diseases. For example, MRI for vascular parkinsonism [3], and levodopa (medication that changes into dopamine in the brain) can suggest PD if they lead to an improvement of the symptoms [8].

The accuracy of PD diagnosis still has room for improvement, with current studies reporting an accuracy of 83.9% when done by experts and 73.8% when assessed by non experts [9]. This can be

due to many reasons, the most common of which are the misdiagnosis of essential tremor, Alzheimer's disease, and vascular parkinsonism [8].

DaTscan and SWEDD

DaTscan is a highly accurate exam, with 98% sensitivity and 100% specificity, in detecting deficiency of dopamine in subjects with parkinsonism [3]. The use of DaTscan for the diagnosis of PD has been studied [10], but it does not add enough to the diagnostic assessment to make it worthwhile [3], since this exam is expensive and not easily available.

However, there are subjects that are clinically diagnosed with PD, whom after post-mortem examination or via DaTscan are determined to not have dopamine deficiency. These subjects are grouped as Scans Without Evidence of Dopamine Deficiency (SWEDD) but it should not be considered a diagnosis, since it can be a mix of different diseases, for instance supranigral parkinsonism and vascular parkinsonism [11].

2.2 Magnetic Resonance Imaging

Magnetic Resonance Imagining (MRI) scanners are used to generate images of different organs of the body with the use of strong magnetic fields, magnetic field gradients, and radio waves. These scans are loud and tend to take a long time to perform, the subjects need to enter a narrow, confining tube where they have to stay still, and they cannot have non-removable metal inside their body. As such, with subjects that are claustrophobic, or due to age or medical conditions cannot lay still, it can be hard to obtain clean images [12].

There are different types of MRI, depending on the settings of the scanners, which can generate different images. Some of the most common types are structural Magnetic Resonance Imaging (sMRI), Diffusion, and Functional. Some of the most common structural images are weighted T1 and weighted T2, though details on these differences are beyond the scope of this project. Different MRI can be used for the same purpose, for instance to identify subjects with PD [13], but for this project only structural Magnetic Resonance Imaging, T1 weighted (sMRI T1w) is used.

sMRI T1w

Figure 2.3 is of a typical MRI, seen from three different directions. In Figure 2.4, it demonstrates how neurons and their components are represented in an sMRI T1w.

In an MRI image, depending on what type it is, different tissues, cells and organs will have a different grey scale, since they are not quantitative, it does not output numbers but instead they report on differences that occur in density. In Table 2.1 it is possible to see what the grey scale values for a sMRI T1w represent, with low values being the darkest ones.

Artifacts due to magnetic fields, noise due to movement, and ghosting (when the same body part shows up multiple times, possibly causing an overlay), are some of the concerns when analysing a MRI.

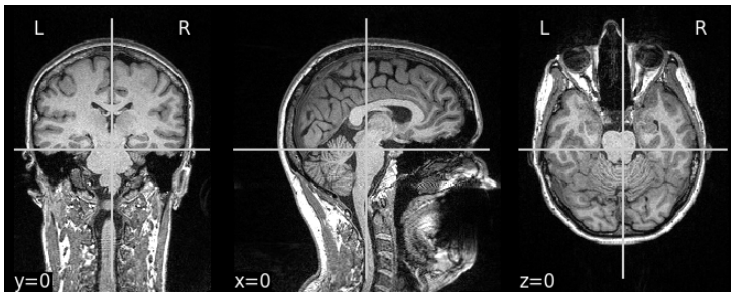


Figure 2.3: Typical MRI

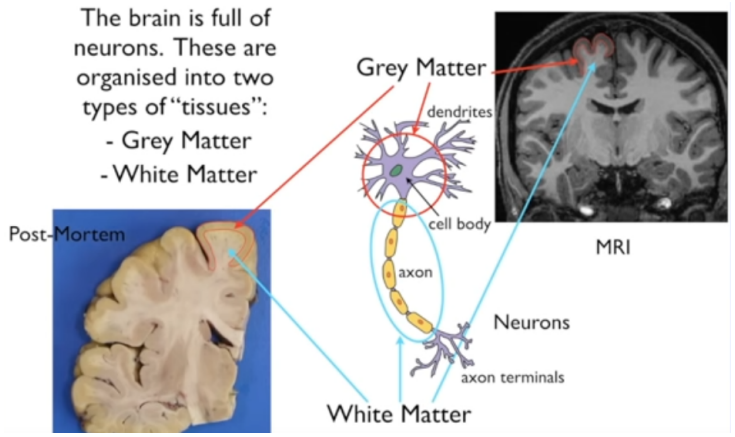


Figure 2.4: Basic anatomy in MRI [14]

Table 2.1: Intensities in sMRI T1w [12]

Signal	**T1-weighted**
High	• Fat • Subacute hemorrhage • Melanin • Protein-rich fluid • Slowly flowing blood • Paramagnetic or diamagnetic substances, such as gadolinium, manganese, copper • Cortical pseudolaminar necrosis • Anatomy
Intermediate	• Gray matter darker than white matter
Low	• Bone • Urine • CSF • Air • More water content, as in edema, tumor, infarction, inflammation, infection, hyperacute or chronic hemorrahage • Low proton density as in calcification

In Figure 2.5 it is possible to see some ghosting in the middle image, since the nose appears on the left and right. In the left image the noise is from lateral movement from the subject, which generates lines across the white and grey matter.

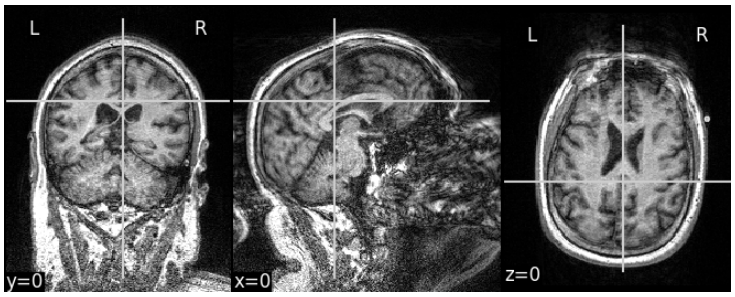


Figure 2.5: Noisy MRI

2.3 Machine Learning Diagnosis

Machine Learning (ML) has been used to distinguish PD subjects from others, through the use of symptoms, speech, movement, and neuroimaging data, refer to systematic reviews such as [15–17]. This section will be focused on the multiple efforts that have been published that use MRI data, some identifying PD by comparison to healthy controls, others by doing differential diagnosis between PD and other diseases, and some detail focus on articles and studies that tried to differentiate PD from SWEDD.

Since the goal is to use MRI data and ML tools to help with the PD diagnosis, particularly in regards to dopamine deficiency, several databases like PubMed, Scopus and Google Scholar were used to find publications that would be relevant to the work.

2.3.1 Machine Learning Pipeline

The typical steps involved in creating a ML model for MRI data, are:

1. **Image processing**, where an MRI image is transformed into a feature vector. Such features can be widely variable. In deep learning models, this step is skipped, since the raw image data is used;

2. **Feature analysis and selection**, with the use of analysis such as Principal Component Analysis, has the goal of downsizing the number of features to improve computation time and remove some of the complexity;

3. **Model training**, where mathematical models are trained in a given dataset, to find the best hyperparameters that achieve the best prediction of the label of each data point. Sometimes, model ensembling and stacking is done to combine different models, with the goal of achieving a higher prediction rate or to avoid overfitting.

4. **Model testing**, which should be performed on a reserved and independent set of data from the data used in previous steps, so as to be able to apply the chosen trained model and to report on its prediction reliability for labels of new, never before seen data. With small amounts of data, reserving a subset of the data might not be possible, and for those cases cross-validation is used.

In using ML as opposed to statistical analysis, the main goal is to achieve prediction, rather than finding group level differences [18]. As such, a big concern that needs to be taken into account is the

use of an independent test set that can be used in the end to test the significance of the results found and its abilities to generalize and predict correctly for new data. Leakage in cross-validation is a concern [18], although when the dataset is small it might be the only option available to be able to proceed with the study [19].

2.3.2 State of the Art

There are multiple reviews done about studies trying to develop algorithms that can be used to diagnose PD. Some reviews comment on the possibility of exploring Deep Learning to accomplish this, but some of the most common critiques are the lack of data quantity and that improvement in the model accuracy does not seem significant compared to the loss of interpretability. Articles like [20], [21], and [22] go into more detail on what has been done, and in Figure 2.6 we can see a comparison between different articles, but this discussion is not in the scope of this project.

TABLE 5 | Overview of papers using deep learning techniques for PD diagnosis.

References	Year	Database	Modality	Method	Modality			Accuracy (%)	
					PD	NC	SWEED	PD/NC	SWEED/NC
Ortiz et al. (2016)	2016	PPMI	SPECT	DNN	–	–	–	95.0	–
Martinez-Murcia et al. (2017)	2017	PPMI	SPECT	3D-CNN	158	111	32	95.5 ± 4.4	82.0 ± 6.8
Choi et al. (2017)	2017	PPMI	SPECT	3D-CNN	431	193	77	96.0	76.5
		SNUH[a]	SPECT		72	10	–	98.8	–
Esmaeilzadeh et al. (2018)	2018	PPMI	sMRI + DI[e]	3D-CNN	452	204	–	1.0	–
Martinez-Murcia et al. (2018)	2018	PPMI	SPECT	DCAE	1,110	195	–	93.3 ± 1.6	–
Sivaranjini and Sujatha (2019)	2019	PPMI	SPECT	2D-CNN	100	82	–	88.9	–
Zhang et al. (2018b)	2018	PPMI	sMRI + DTI	GCNN	596	158	–	95.37 (AUC)	–
McDaniel and Quinn (2019)	2019	PPMI	sMRI + DTI	GCNN	117	30	–	92.14	–
Shen et al. (2019b)	2019	HSHU[b]	PET	DBN	100	200	–	90.0	–
		WXH[c]	PET		25	25	–	86.0	–
Shen et al. (2019a)	2019	Multi-site[d]	TCS	DPN	76	77	–	86.95 ± 3.15	–

[a]*SNUH, Seoul National University Hospital cohort.* [b]*HSH, HuaShan Hospital cohort.* [c]*WXH, WuXi 904 Hospital cohort.* [d]*Shanghai East Hospital of Tongji University and the Second Affiliated Hospital of Soochow University.* [e]*DI, Demographic Information.*

Figure 2.6: Comparison table from review [20], for PD diagnosis with Deep Learning.

There is a large quantity of articles related to diagnosis of PD through the use of more classic ML. However, there is no standard on how to report the results and what information is essential, and the studies most of the time lack transparency in how they achieved the results and what steps and decisions were made. Moreover the reviews that compile these studies do not go into detail and do not report all the methods used in the different studies, making it difficult to do a fair comparison.

In this section comparisons done by review articles are shown in Table 2.2 and Figures 2.7, 2.8, 2.9, and 2.10.

The validation or testing method chosen to obtain the results is not mentioned, for example Table 2.2, which can limit the understanding how some accuracies were achieved. Some results, like 100% accuracy are very unexpected to obtain in a ML model with a proper independent test, especially when dealing with data as complex as neural imaging.

Table 2.2: Comparison table from review [18], for PD diagnosis.

Ref	Groups (N)	Method	Input Modalities	Accuracy
(Focke et al., 2011)	Controls (22) - PD (21) Controls (22) - PSP (10) Controls (22) - MSA (11) MSA (11) - PD (21) MSA (11) - PSP (10) PD (21) - PSP (10)	SVM (linear)	T1	42 % 93.7 % 78.8 % 71.9 % 76.2 % 96.8 %
(Cherubini et al., 2014)	PD (57) - PSP (21)	SVM (kernel not specified)	T1, T2, DTI	100 %
(Skidmore et al., 2015)	Controls (22) - PD (20)	Bootstrap	DTI	90.1 %
(Marquand et al., 2013)	PSP (17), PD (14), MSA (19) Controls (19), PSP (17), PD (14), MSA (19) PSP (17), PD (14), MSA-C (7), MSA-P (12) Controls (19), PSP (17), PD (14), MSA-C (7), MSA-P (12)	Multinomial logit	T1	91.7 % 73.6 % 84.5 % 66.2 %
(Filippone et al., 2012)	Controls (14), PD (14), PSP (16), MSA (18)	Multinomial logit	T1, T2, DTI	75.3 %
(Salvatore et al., 2014)	Controls (28) - PD (28) Controls (28) - PSP (28) PD (28) - PSP (28)	SVM (linear kernel)	T1	92.7 % 97 % 98.2 %
(Duchesne et al., 2009)	PD (16) - PSP (8) + MSA (8)	SVM	T1	90.6 %
(Haller et al., 2012)	PD (17) - Other (23)	SVM (Gaussian kernel)	DTI	97.5 %
(Haller et al., 2013)	PD (16) - Other (20)	SVM (Gaussian kernel)	SWI	86.9 %

Table 1. Summary of Studies Using Magnetic Resonance Imaging and Artificial Intelligence in Studies of Diagnosis of Parkinson's Disease[a]

papers	MRI modalities	data source	subjects	feature extraction method	validation method	classifier	acc (%)	affected brain regions
38	rs-fMRI	recruited	51 PD, 50 NC	group difference analysis	LOOCV	SVM	84.2	lingual gyrus, putamen, cerebellum posterior lobe
39	multimodality: rs-fMRI, sMRI	recruited	19 PD, 27 NC	group difference analysis	LOOCV	SVM	86.96	ORBmid, ROL, PHG, ANG, MTG, PCL, PreCG, PCG
40	sMRI	PPMI	9 PD, 6 NC	scale invariant feature transform	LOOCV	SVM	80	limbic lobe, frontal lobe, sublobar, midbrain, pons, posterior lobe, occipital lobe
41	rsfMRI	recruited	21 PD, 26 NC	Kendall tau rank correlation coefficient	LOOCV	SVM	93.62	DMN, control network, cerebellum, etc.
42	sMRI	PPMI	374 PD, 169 NC	joint feature-sample selection	10-fold cross-validation	LDA	81.9	red nucleus, substantial nigra, pons, middle frontal gyrus, superior temporal gyrus
43	sMRI	PPMI	369 PD, 169 NC	joint kernel-based feature selection	10-fold cross-validation	SVM	70.5	insula, cingulate gyrus, hippocampus, parahippocampal gyrus, amygdala, caudate, putamen, etc.
44	rsfMRI	recruited	80 PD, 84 NC	previous quantitative meta-analyses	10-fold cross-validation	SVM	75	rsfMRI networks subserving autobiographical or semantic memory, motor execution, and theory-of-mind cognition
45	sMRI	PPMI	69 PD, 103 NC	filter- and wrapper- based features extraction	10-fold cross-validation	SVM	85.78	frontal lobe, parental lobe, limbic lobe, temporal lobe, and central region
46	sMRI	PPMI	263 PD, 123 NC	the R-package CARET	5-fold cross-validation	SVM, AdaBoost	>80	superior parietal gyrus, putamen, caudate
47	sMRI	PPMI	374 PD, 169 NC	random forests	10-fold cross-validation	SVM	93	frontal, occipital, and temporal lobes, limbic lobe, brainstem, midbrain
48	multimodality: sMRI, R2*, DTI	recruited	34 PD, 31 NC	FCP, Lasso, PCA	bootstrapping	FCP	99.7	amygdala, caudate, substantia nigra, pallidus, hippocampus, red nucleus, dentate
49	sMRI	recruited	45 PD, 40 NC	recursive feature elimination	cross-validation	SVM	>95	the cerebellar Crus I

[a]Abbreviations: acc, accuracy; SVM, support vector machine; PD, Parkinson's disease; NC, normal controls; LOOCV, leave-one-out cross-validation method; rsfMRI, resting state functional MRI; sMRI, structural MRI; ORBmid, middle frontal gyrus, orbital part; ROL, Rolandic operculum; PHG, parahippocampal gyrus; ANG, angular gyrus; MTG, middle temporal gyrus; PCL, paracentral lobule; PreCG, precentral gyrus; PCG, posterior cingulate gyrus; DMN, default mode network; LDA, linear discriminant analysis; FCP, folded concave penalized; PCA, principle component analysis-based feature selection.

Figure 2.7: Comparison table from review [19], for PD diagnosis.

Accuracy is used in every comparison, but it can be misleading due to unbalanced datasets. Paper 43 in Figure 2.7 is an example of this: since the dataset is unbalanced, a model that would output every subject with PD would have an accuracy of 68.6%, which is very close to the accuracy reported.

The MRI modality is reported, but the feature extraction method and the type of features used are not shown in the comparisons, and in the articles are not explained in great detail. This makes comparison between different papers more difficult, since with different features for the same MRI modality, different results may be able to be achieved.

The most common models reported to be used in several of the reviews are SVM, with some mentions of Random Forest, Naive Bayes, and Logistic Regression. The accuracy in distinguishing between healthy controls and PD subjects varies between 80% and 100%. Moreover, distinguishing between healthy controls and different diseases tends to yield higher accurate models, then when trying to differentiate diagnosis [18].

Table 2. Summary of Studies Using Magnetic Resonance Imaging and Artificial Intelligence in Studies of Differential Diagnosis of Parkinson's Disease[a]

papers	MRI modalities	data source	subjects	feature extraction method	validation method	classifier	acc (%)	affected brain regions
62	sMRI	recruited	28 PD, 28 PSP	PCA	LOOCV	SVM	88.9	midbrain, pons, corpus callosum, and thalamus
63	sMRI	recruited	21 IPS, 11 MSA, 10 PSP, 22 NC	group comparison analysis	LOOCV	SVM	IPS vs PSP 96.8 IPS vs MSA 71.9	putamen, superior parietal lobe, precuneus, external capsule, corticospinal tract, precentral gyrus, occipital pole, pons, mesencephalon, dorsal basal ganglia, cerebellar peduncles
64	sMRI	recruited	204 PD, 106 PSP, 21 MSA-C, 60 MSA-P, 73 NC	group comparison analysis	LOOCV	SVM	>80	midbrain, basal ganglia, and cerebellar peduncles
65	SWI	recruited	16 PD, 20 other Parkinsonism	ReliefF feature selection	10-fold cross validation	SVM	86	thalamus and substantia nigra
66	DTI	recruited	17 PD, 23 other Parkinsonism	ReliefF feature selection	10-fold cross validation	SVM	97	bilateral network, predominantly in the right frontal white matter
67	multimodality: DTI, sMRI	recruited	21 PSP, 57 PD	group comparison analysis	LOOCV	SVM	100	basal ganglia, midbrain, cerebellum, corpus callosum
68	multimodality: DTI, sMRI	recruited	15 rET, 15 tPD	group comparison analysis	LOOCV	SVM	100	caudate nucleus, globus pallidus, midbrain, internal capsule, body of the corpus callosum, and cerebellum
69	sMRI	recruited	14 PDD, 15 PDMCI, 16 PDCI	FSS	5-fold cross validation	naive Bayes	PDD vs PDCI 93 PDD vs PDMC 96 PDMCI vs PDCI 86 PDD vs PDMCI vs PDCI 64	caudate, entorhinal cortical, hippocampus, brain stem, cerebellum, lateral ventricle
70	multimodality: DTI, sMRI	recruited	12 left -sided and 12 right-sided symptom onset PD	FSS	LOOCV	SVM	96	right hippocampus

[a]Abbreviations: acc, accuracy; SVM, support vector machine; PD, Parkinson's disease; PSP, progressive supranuclear palsy; MSA, multiple system atrophy; IPS, idiopathic Parkinson syndrome; NC, normal controls; LOOCV, leave-one-out cross-validation method; rsfMRI, resting state functional MRI; sMRI, structural MRI; SWI, susceptibility-weighted imaging; DTI, diffusion tensor imaging; PCA, principle component analysis-based feature selection; rET, tremor with rest tremor; tPD, tremor-dominant Parkinson's disease; PDD, Parkinson's disease with dementia; PDMCI, Parkinson's disease with mild cognitive impairment; PDCI, Parkinson's disease cognitively intact; FSS, Feature subset selection; MSA-C: MSA patients with a cerebellar syndrome, MSA-P: MSA patients with a parkinsonian type.

Figure 2.8: Comparison table from review [19], for differential diagnosis.

Name of author and year	Modality	Segmentation and classification techniques used	Brain disorder	Overall accuracy	Size of dataset
Adeli et al. (2016)	T1 weighted	Region of interest volume features with LS- LDA	Schizophrenia dieses Parkinson's disease	82%	Large Dataset consist of T1 weighted MRI scans of 374 Parkinson's disease patients and 169 normal control patients (PPMI) (https://www.ppmi-info.org/access-data-specimens/download-data/)
Peng et al. (2017)	T1 weighted	Filter- and wrapper- based feature selection methods with multi-kernel support vector machine (SVM)	Parkinson's disease	86%	Dataset consist of T1 weighted MRI scans of 69 Parkinson's disease patients and 103 Healthy control patients (PPMI)(https://www.ppmi-info.org/access-data-specimens/download-data/)
Amoroso et al. (2018)	T1 weighted	Features from the Network measures and clinical scores with Random Forest + Support Vector Machine	Parkinson's disease	93%	Dataset consist of T1 weighted MRI scans of 374 Parkinson's disease patients and 169 Healthy control patients (PPMI) (https://www.ppmi-info.org/access-data-specimens/download-data/)
Oliveira et al. (2017)	I^{123}-CIT SPECT (single-photon emission computed tomography)	SVM	Parkinson's disease	98%	Dataset consist of I^{123}-CIT SPECT scans of 443 Parkinson's disease patients and 209 Healthy control patients (PPMI) (https://www.ppmi-info.org/access-data-specimens/download-data/)

Figure 2.9: Comparison table from review [23], for PD diagnosis.

Table 3 Comparison of different methods based on machine learning and deep learning for the identification and classification of brain diseases

Method	Type of method	Imaging technique	Brain disease detection	Reference	Dataset
Kernel SVM	ML	T_{s1}MRI	Parkinson Disease	Wang S [87]	Dataset was obtained from the patients affected from Parkinson disease and normal healthy person
Random forest and SVM	ML	T_{s1}MRI	Parkinson disease	La Rocca M et al. [88]	Dataset was obtained from the patients affected from Parkinson disease and normal healthy person
SVM, KNN, and logistic regression	ML	MRI	Parkinson disease	Faria DB [89]	Dataset was obtained from the patients affected from Parkinson disease and normal healthy person

Figure 2.10: Comparison table from review [24], for PD diagnosis.

PD vs SWEDD

In this section only for articles that studied similar hypothesis of the one for this thesis are compared. As such, studies referencing exams and tests other than MRI were not to be considered. In addition, articles using sMRI T1w are more relevant to the work intended, as well as the ones using classic ML algorithms instead of deep learning.

With this objective, a search was done in PubMed and Scopus using the query `((MRI) OR (T1) OR (DTI) OR (Magnetic Resonance) OR (Diffusion)) AND ((DaTscan) OR ("DaT SPECT") OR (SWEDD) OR ("Dopaminergic Deficit") ) AND ((ML) OR (Machine learning) OR (AI) OR (Artificial Intelligence) OR (Classif*) OR (Predict*) OR (DL) OR (Deep Learning))`

This query returned very few studies groups by dopamine (PD, SWEDD, and control groups) with MRI and ML, returned very few studies, see Table 2.3. Of these results only the one uses sMRI T1w. The others use DTI images and classical ML algorithms with features extracted from the images (DTI tracts), reporting an accuracy of up to 97%.

Therefore, work has yet to be done in order to answer the hypothesis this thesis raises: whether it is possible to detect dopamine deficiency with ML and sMRI T1w, in order to assist in PD diagnosis.

Table 2.3: Comparison table of studies that differentiate PD, Control and SWEDD

Reference	MRI type	Features	Model	Sample Size	PD	Control	SWEDD	Test	Accuracy
[25]	DTI	DTI tracts	SVM	142	37%	37%	27%	42%	81.25%
[26]	DTI	DTI tracts	SVM quadratic	48	54%	-	46%	LOOCV	72.5%
[27]	DTI	DTI tracts	SVM linear	80	33%	33%	33%	LOOCV	77.92%
[28]	DTI	DTI tracts	SVM	77	48%	-	52%	LOOCV	97%
[29]	sMRI T1w	Image slices	CNN	197	43%	43%	15%	30%	60%-80%

The last reference from the Table 2.3, is the dissertation from one of the students who is part of the lab with which the work of this thesis was developed. This is also the only study found that uses sMRI T1w to distinguish PD from SWEDD with ML. A 2vs2 group approach was employed (PD vs Control, PD vs SWEDD, SWEDD vs Control), always with 50% of each group, and the results can be seen in Figure 2.11. The orange plots that are resulting from the use of SPECT images may be a sign of a promising prospect, but these results are outside of the scope of the thesis.

Additionally, it was reported to predict with an accuracy of 97.4%, 73.3% and 65.3% in separating PD vs Control, PD vs SWEDD and Control vs SWEDD, respectively, by choosing adequate MRI slices. In particular, this suggests, that Control and SWEDD groups might not be not easily distinguishable.

2.4 Principal Component Analysis

2.4.1 PCA

Principal Component Analysis (PCA) [30] is a process that takes a dataset with multiple real features and computes its ”principal components”, which are unit vectors in the feature space. These are created sequentially, and each principal component is the linear combination of features that explains the maximum amount of variance possible while being orthogonal to the principal components extracted before.

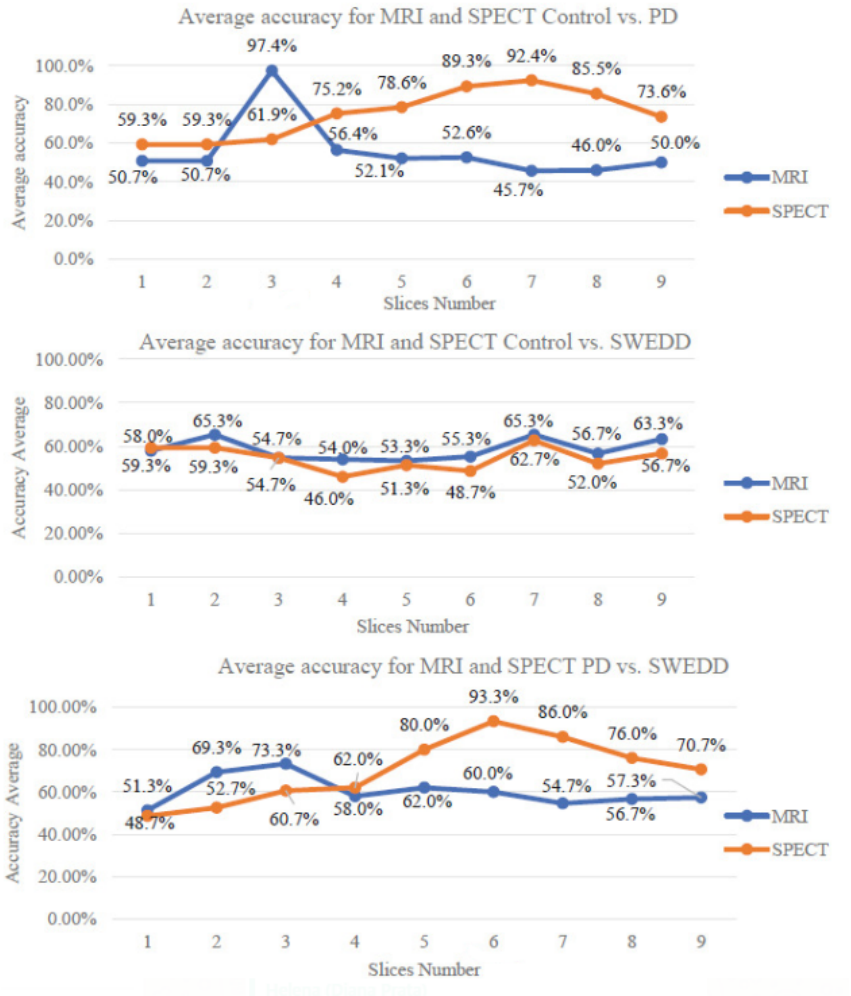


Figure 2.11: Accuracy by using different MRI slices [29]

In addition, the method computes how much variance each principal component explains.

PCA is commonly used for dimensionality reduction. By using the principal components that explain the most variance as features, it is possible to transform the dataset and reduce the number of features while ideally preserving as much useful information as possible.

The most common methods employed to find the principal components use matrix factorization techniques. It must be noted that the number of principal components is always limited by both the total number of features and the number of datapoints. As we will see in Section 4.3.1, PCA will be applied to a dataset with more features than subjects, therefore the total number of principal components equals the number of datapoints.

2.4.2 RPCA

As we will see in section 4.3.1, PCA can be affected by the presence of outliers in the dataset. To correct this, another method called Robust Principal Component Analysis (RPCA) was proposed in the literature [31]. RPCA works in largely the same way as PCA, but the principal components it outputs are calculated in a different way that is more robust to outliers. The implementation of this method that was

used is from this source [32], where the original articles are cited and explained.

2.5 Machine Learning Algorithms

In this section, the ML algorithms are described alongside what are some of the more important hyperparameters to be chosen and how they affect the algorithm.

2.5.1 Random Forest

Random forests [33] are an ensemble ML method that can be used for classification. The idea is to train multiple decision trees for classification using different subsets of the same training data, and define the output of the model as a whole as the most common output among all trees. This is done because a single tree tends to overfit the training data and thus lead to a model with high variance, therefore training multiple trees provides a way of reducing the variance of the model.

To construct a decision tree, a subset of the training dataset is first extracted. Then, a tree is built top-down by choosing at each step the feature that better separates the corresponding set of points. Each note represents a decision according to a feature, and the two edges leaving it represent the two possible answers to said decision. This is done until either all features are used or it is not worth to further separate the data.

At each leaf, the probability of the datapoint belonging to each class is computed. When computing the output of the random forest, the probabilistic output of each tree is averaged out.

2.5.2 Support Vector Machines (SVM)

Support Vector Machines [34] are supervised learning models used for data classification. Given a set of datapoints with each one assigned to one of two classes, an SVM can be employed to train on such data and define a binary classifier capable of labeling new datapoints.

Given a training dataset of points with N features, the simplest application of a SVM is to view them as a set of N dimensional vectors and determine a $(N-1)$-dimensional hyperplane which better separates the points. This defines what is called a linear classifier.

To create more sophisticated classifiers, a transformation of the feature space can be employed, possibly mapping the data points to higher-dimensional spaces. If the transformation is non linear, the classifier may be non-linear in the original feature space. The choice of the transformation may make the learning process difficult or even unfeasible. However, if such a transformation is carefully chosen, the computation can be done rather quickly.

Instead of choosing a transformation directly, it is possible choose a kernel function which is related to the transformation. This function determines the shape of the classifier of the SVM. For non-linear classifiers, the most common choice for a kernel is the Radial Basis Function (RBF).

When training a SVM, multiple hyperparameters can be tweaked to obtain a better result. One parameter used across all SVM kernels is $C > 0$, and acts as the regularization parameter: a low C

makes the decision function simpler at the cost of training accuracy, whereas a high C allows for a more complex decision function which can correctly label more training examples. Associated to the RBF kernel in particular is another parameter $\gamma > 0$: intuitively, the smaller this value is, the further the influence of a sample in labeling new datapoints in its vicinity.

2.5.3 Logistic Regression

Logistic Regression [35] is a model used to estimate the probability of a datapoint belonging to one of two classes. Using this model, such probabilities are modeled using a logistic function whose parameters are learned through training.

When training a Logistic Regression, one hyperparameter that can be tweaked to obtain a better result is C, which is a regularization parameter, like in a SVM.

2.5.4 Perceptron

The Perceptron [36] is a linear classifier used to separate a set of datapoints into two classes. This is achieved by employing a linear function with the features as inputs. The model labels the datapoint depending on whether the output of the function is above a certain threshold or not.

When training a Perceptron, a hyperparameter that can be changed to obtain a better result is $alpha$ which is a penalty parameter, and it is equivalent to $1/C$ where the C is the regularization parameter in SVM.

2.5.5 Ridge Classifier

The Ridge Classifier [37] is a binary classifier based on the Ridge Regression.

Given a set of datapoints, each with N features and an additional target value, a linear model attempts to compute the target value using some linear combination of the features. Ridge Regression finds the coefficients of the model by minimizing the error function from Ordinary Least Squares plus a regularization term which imposes a penalty on the size of the coefficients. To use this model for classification, the target values are converted into $\{-1, 1\}$ depending on the class each datapoint falls into, and then the same methodology is applied.

When training a Ridge Classifier, a hyperparameter that can be adjusted to obtain a better result is $alpha$ which is the same as in a Perceptron.

Chapter Summary

This chapter introduced Parkinson's Disease along with background on its diagnosis process and how other diseases can be related. Then an example of a typical pipeline for applying ML on these images was presented.

Additionally an overview of relevant literature, with more detail on papers that studied similar hypothesis to the one proposed by this thesis was given and commented on. Finally, Principal Component Analysis and ML algorithms were introduced.

Chapter 3

Methods

This chapter begins with an overview of the available data and the decisions for selecting it. This is followed by an explanation of the processing performed on the images to extract features and what these features are. Additionally, the approaches that were planned to be used with this data are introduced.

3.1 Data Source

The sMRI T1w comes from the Parkinson's Progression Markers Initiative (PPMI) database [38]. PPMI is a landmark observational study that makes its data set and biorepository available to academia and industry. The PPMI study divides its enrolled subjects into different groups, called *Research Groups*:

- **PD** - Subjects with a PD diagnosis for two years or less that are not taking PD medications;
- **Control** - Subjects without PD who are 30 years or older and who do not have a first degree blood relative with PD;
- **SWEDD** - Subjects that enrolled as PD subjects who have DaTscans that do not show evidence of a dopaminergic deficit;
- **GenCohort PD** - Subjects with and without PD who have a genetic mutation in LRRK2, GBA, or SNCA;
- **Gen Reg PD** - Subjects with and without PD who have a genetic mutation in LRRK2, GBA, or SNCA or a first-degree relative with a LRRK2, GBA, or SNCA mutation who are evaluated at less frequent intervals to augment and broaden the follow-up of PD subjects and family members with PD associated mutations.

3.2 Image Selection

For each subject only one 3D image was selected, to avoid subject-bias in the hopes of achieving the best classification, facilitating the interpretation of results and gathering of insights. In the PPMI database

there are 761 subjects with an MRI and DaTscan. From this group various choices in parameters and restrictions were chosen:

1. MRI had to have *Field Strength* = 3T, left 471 subjects;
2. MRI had to be *3D*, left 470 subjects;
3. MRI had to be *MPRAGE*, left 407 subjects;
4. MRI without considerable noise or ghosting, left 402 subjects;
5. MRI had to have *Repetition Time* = 2300, left 337 subjects;
6. MRI and DaTscan within 12 months, left 316 subjects

So, each subject had to have a DaTscan result and an sMRI T1w, within one year so as to be as close to the ground truth as possible. Furthermore, the sMRI T1w had to have good quality and minimal noise and ghosting, selected via a manual visual inspection. In Figure 3.1 there is an example of a MRI that is to be excluded. Additionally it had to have the following parameters: *MPRAGE*, *Field Strength*=3T, *Repetition Time (TR)*=2300. These parameters were chosen to avoid possible bias that could exist due to relations between the different parameters and labels.

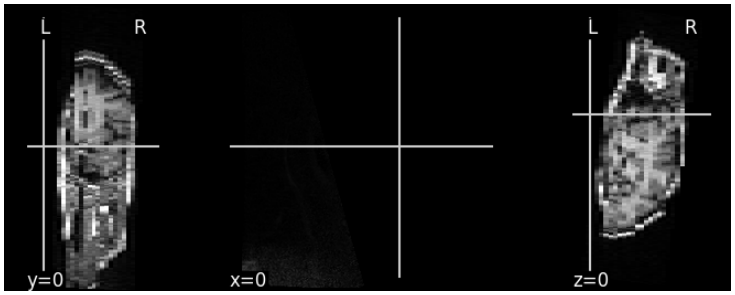


Figure 3.1: MRI that is to be excluded

There were 5 subjects whose DaTscan result did not match what would be expected from belonging to the *Research Group*, and these were excluded. Once the model is chosen and trained, they may be used to check what the model would classify them as.

Chosen Data

From the 311 chosen subjects, 104 (33%) had a negative DaTscan, and 207 (67%) a positive DaTscan. Of those with negative label, 71 were Control and 33 SWEDD, while of those with a positive label, 160 were PD, 46 were GenCohort PD, and 1 was GenReg PD.

3.3 Image Preprocessing

There are multiple resources that can be used to process MRIs, the purpose of which is to turn these 3D images into features that can be calculated. These can be for example surface areas, folding indexes, or volumes of regions of the brain.

FreeSurfer [39] provides a full processing stream for sMRI data [40], including: skull stripping (Figure 3.2), gray-white matter segmentation, reconstruction of cortical surface models, labeling of regions on the cortical surface, as well as subcortical brain structures (Figure 3.3). This is the software used by Diana Prata's Lab, and Vasco's pipeline, which was used as a baseline for this project.

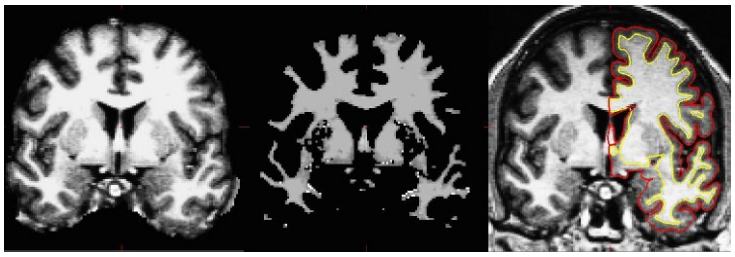


Figure 3.2: Three stages from the FreeSurfer cortical analysis pipeline: A - Skull stripped image. B - White matter segmentation. C - Surface between white and gray (yellow line) and between gray and pia (red line) overlaid on the original volume [39]

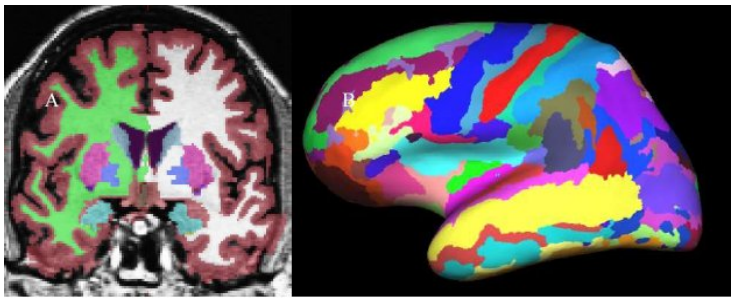


Figure 3.3: A - Volume-based labeling. Note that cortical gray matter and white mater are represented by single classes. Also note that there are separate labels for the structures in each hemisphere. B- Surface-based labeling. [39]

3.3.1 Resulting Features

This processing turns an sMRI T1w image into 689 features, 8 of which are copies or masks and can be removed. The remaining are separated between left and right hemisphere (*lh*, *rh*), and these features can be divided between 41 different brain regions: *bankssts*, *basalganglia*, *brain*, *caudalanteriorcingulate*, *caudalmiddlefrontal*, *cerebellum*, *corpuscallosum*, *cuneus*, *entorhinal*, *frontalpole*, *fusiform*, *hippocampus*, *inferiorparietal*, *inferiortemporal*, *insula*, *isthmuscingulate*, *lateraloccipital*, *lateralorbitofrontal*, *lingual*, *medialorbitofrontal*, *middletemporal*, *paracentral*, *parahippocampal*, *parsopercularis*, *parsorbitalis*, *parstriangularis*, *pericalcarine*, *postcentral*, *posteriorcingulate*, *precentral*, *precuneus*, *rostralanteriorcingulate*, *rostralmiddlefrontal*, *superiorfrontal*, *superiorparietal*, *superiortemporal*, *supramarginal*, *temporalpole*, *thalamus*, *transversetemporal*, *ventricle*.

These features can be of 9 different types: *thickness*, *volume*, *curvind*, *wm*, *thicknessstd*, *foldind*, *gauscurv*, *meancurv*, *area*.

3.4 Correlations

Since the number of features was higher than the number of subjects, this was a situation that normally leads to poor results in Machine Learning. As such feature selection was important for this case.

Correlation analysis can identify which features are very similar, and by choosing a threshold, such features can be eliminated which would not significantly contribute to the model choice.

3.5 Training, Validation, and Testing Sets

An independent test set is important for testing the validity of the results obtained and to test the final model as a possible tool to be used in the diagnosis process. Therefore, a representative and significant test set was important, so 30% of the data was reserved for final testing. Some considerations regarding the diagnosis, sex and age, had to be accounted for, since the amount of data was not large enough to rely on randomness as a guaranteed tool to achieve representation of all classes.

The age of the subjects has to be balanced, since it can be a confounding variable in MRI images, because age has an effect on the brain. For instance, it is negatively correlated with grey matter measures [41]. This was accomplished by ensuring the average age is the same in both sexes. Furthermore, the different diagnosis in each DaTscan label was also balanced such that both sexes have similar percentages of each diagnosis, and subjects with multiple images were left to the test set, since they could bring a higher value in the verification of the reliability of the results. Due to there being a relationship between sex and PD diagnosis, as there is a higher percentage of men diagnosed [42], the balancing in the test set was done such that it was equivalent across diagnoses, with 1 female to every 1.5 male, so that the confidence estimates could be equal for both sexes in the test results analysis.

So, after imposing these restrictions, subjects were chosen randomly from the remaining options, to be either in the training or testing sets.

Validation

To compare different approaches, either cross-validation can be used, or a validation set. For the creation of a validation set, 10% of the training set was selected randomly, and for the cross-validation all of the training set was used.

3.6 Data Transformations

Different transformations could be made on the data, for efforts of feature selection or to remove subject specific data to make comparing data between subjects more fair.

3.6.1 Brain Regions

The 672 features can be divided into 41 sets related to certain brain regions. This transformation allows there to be fewer features than subjects, and it can also be a method to identify which brain regions are more important for the classification problem.

This was a new idea proposed by this thesis, to use *a priori* knowledge that existed of the features and data being used, instead of relying on a naive search for feature selection.

3.6.2 RPCA

Features obtained by explaining 90% of variance with RPCA can be used instead of the original 672. By using this transformation, 50 features can be used which transforms the problem into one where there are fewer features than subjects.

One possible problem with this transformation was that since the number of subjects was less than the number of original features, the new features explain not just variance of features but also of subjects. As a result this features might not have been well chosen for generability and future data.

3.6.3 Normalize

The normalization of data is a transformation that is performed before RPCA, but using it by itself can improve some models, such as SVM, since it allows the models to not give more significance to certain features due to their having higher values.

This transformation is done such that for each feature

$$x' = \frac{x - \bar{x}}{std}$$

, where std is the standard deviation, and $\bar{x}$ is the mean.

3.6.4 Relative

The features which are of type area or volume, can be divided by the total surface area of the brain, or total volume of the brain, so as to have features that can be comparable between subjects.

3.7 Balancing

This dataset was unbalanced, and most ML algorithms perform better when training with balanced data. As such multiple techniques for balancing the data are explored.

3.7.1 Balancing in model

Most models implemented by *sklearn*, and in particular all models used for this thesis, have a parameter that can be used so the model gives weights to the classes, and for balancing, this parameter

can be set to balanced, `class_weight='balanced'`.

3.7.2 Undersampling

Another option is to perform undersampling, which consists of removing samples from the class with the highest amount, so as to balance them.

3.7.3 Oversampling

Another option is to perform overrsampling, which consists of repeating samples from the class with the least amount, so as to balance them.

3.8 Baseline Approach

Vasco Sá, a student from the same research laboratory with which this project was produced, had previously developed an MRI ML pipeline to create and train a model to diagnose Alzheimer's disease, and those results are currently in the process of being published as an article.

This approach consists of using a voting system between 7 different classifiers: a linear support vector machine (l-SVM); a decision tree classifier (DT); a random forest classifier (RF); an extremely randomized tree classifier (ET); a linear discriminant analysis classifier (LDA); a logistic regression classifier (LR); and a logistic regression classifier with stochastic gradient descent learning (LR-SGD).

All hyperparameters were chosen by using an evolution algorithm and cross validation.

This pipeline serves as a baseline approach.

3.9 Exploration for best model

The main idea was to test simple models, since we wanted to avoid overfitting, which happened in the baseline approach. A possible reason was due to the complexity. Comparing multiple combinations of algorithms, balancing and transformation seemed to be the best approach as, in this way, it was possible to choose the best ones to test in the reserved set.

Here, the multiple options were obtained using combinations of alternatives. Each combination was composed choosing one of each alternative in the following list:

1. Algorithm: Logistic Regression, Perceptron, Ridge Classifier, Random Forest, Support Vector Machine
2. Validation: 10% of training set, Cross-validation
3. Balancing: None, Over, Under, Balanced in model
4. Features: All features, the subset of features for each of the 41 brain regions
5. Transformation: None, Normalize, Relative, RPCA (only when all features are selected)

For the relevant models, cross-validation was used to find the best hyperparameters (discussed in section 2.5), with grid search.

Besides the balancing types mentioned, types which create artificial data were not considered due to the medical nature of the data. Moreover, when using cross-validation and oversampling simultaneously it could have been the case that the training subsets in the folds could have been the same data points as in the testing fold, since *sklearn* was not prepared to handle this situation. Thus, this method was programmed from scratch.

The RPCA transformation was only applied when using all features, since this and selecting a subset of features from a brain region are different feature selection methods.

When using cross-validation, the standard deviation was computed. When validation (10%) was used, the confusion matrix was presented.

In sum, there were a total of 160 combinations if all features were selected, and 4920 when the subset of features corresponds to each of the 41 brain regions.

For each combination the following metrics were calculated: Accuracy, Balanced accuracy, F1-score, Balanced F1-score, and ROC AUC.

3.10 Interpretability

The features obtained through RPCA can be used to calculate how much each original feature was of relevance to the results of a given model. It would be possible to see if the variance explained was related to the importance of that feature.

Since the models chosen were simple, it was possible to extract feature importance after training. For example, the weights associated with each feature in SVM could be calculated, or for the case of a Random Forest, how much each feature decreases the impurity of splits.

By using the subsets of features related to brain regions, it was possible to check which ones gave better results and infer what the more important regions for the identification of dopamine deficiency were.

Since no approach resulted in positive results, as we will see in the next chapters, these explorations were not made.

3.11 Generalizability

To test how well the final chosen model would generalize for new data and test its reliance, multiple options were considered.

Firstly, for the subjects who had multiple images taken on the same day, and which were selected for testing, both images would be passed through the model to check if they model gave the same result. Moreover, other subjects who had multiple images but from different months could also be tested, but since time can affect the deficiency of dopamine, the results might not be equal due to this and not to the unreliable of the model.

Another approach to test how the model would be able to handle new data, would be to test images that have different parameters, such as the slice thickness. If the model gives good results when using other parameters, it could be an indication that in future studies there can be less restrictions on the images parameters and it would more easily be able to be used in a clinical setting.

Finally, for the subjects whose research group and DaTscan result did not match, their images would go through the model to check what result it would give.

Chapter Summary

This chapter gave an overview of the source of the data and explained and how the MRI were chosen, and its extracted features. Then, how the data was separated for validation and testing. Moreover, the data transformations and balancing types that are used in later sections were explained.

Finally, the baseline approach and the approach used for exploring different options were detailed. Along with some final considerations and ideas that were to be used if the hypothesis was proven right.

Chapter 4

Results

This chapter shows the results from analysis performed on the data and the results from the ML models used and tested.

The dataset consisted of 311 subjects and 672 features, with a binary label.

4.1 Dataset Analysis Results

4.1.1 Label Distribution

From the dataset, 33% have a negative DaTscan and 67% have a positive DaTscan. Both training and testing set have the same distribution.

4.1.2 Research Groups Distribution

Subjects who have a negative DaTscan belong to either the Control group or SWEDD group, while subjects who have a positive DaTscan belong to either the PD group or GenCohort PD group or GenReg PD group. Since there was a high percentage of subjects from the GenCohort PD group who had multiple images in the same day, and these were to be reserved for the test set, the distributions of training and testing sets were not equal. The final distributions in both sets can be seen in Table 4.1.

Table 4.1: Distribution of research groups, in the training and testing sets.

	Control	**SWEDD**	**PD**	**GenCohort PD**	**GenReg PD**
Train	23%	11%	53%	13%	1%
Test	23%	11%	47%	19%	0%

4.1.3 Age Distribution

The mean and standard deviation (std) age across all the data is 61.3 $\pm$ 10.1. To test the significance of age in the DaTscan result, the Mann-Whitney test was used, which resulted in a $p = 0.163$, so there is not significant evidence to reject that the age distribution is equal for both labels.

Table 4.2: Age mean and std across label, in the train and test sets.

	DaTscan negative	DaTscan positive
Train	59.1 $\pm$ 11.7	61.8 $\pm$ 9.4
Test	61.8 $\pm$ 10.0	62.5 $\pm$ 9.4

Table 4.3: Age mean and std across research group, in the train and test sets.

	Control	SWEDD	PD	GenCohort PD	GenReg PD
Train	58.2$\pm$ 12.0	61.1$\pm$ 10.8	61.5$\pm$ 9.0	62.6$\pm$ 11.0	70.9
Test	62.0$\pm$ 9.6	61.4$\pm$ 11.3	60.6$\pm$ 9.8	67.2$\pm$ 6.5	-

4.1.4 Sex Distribution

The distribution of the sexes across all the data is 36% female and 64% male. To test the significance of sex in the DaTscan result, the Pearson Chi-Square test was used, which resulted in a $p = 0.762$, so there is not significant evidence to reject that the sex distribution is equal for both labels.

Table 4.4: Sex distribution Female - Male, across label, in the train and test sets.

	DaTscan negative	DaTscan positive
Train	36% - 64%	34% - 66%
Test	41% - 59%	40% - 60%

Table 4.5: Sex distribution Female - Male, across research group, in the train and test sets.

	Control	SWEDD	PD	GenCohort PD	GenReg PD
Train	35% - 65%	39% - 61%	35% - 65%	29% - 71%	-
Test	41% - 59%	40% - 60%	39% - 61%	44% - 56%	100% - 0%

4.2 Correlation Analysis Results

By analysing the number of zero entries in the features, 4 features were found to only have value zero: *Right-non-WM-hypointensities*, *Left-non-WM-hypointensities*, *Right-WM-hypointensities* and *Left-WM-hypointensities*, and they were removed. Moreover, the features *5th-Ventricle* and *non-WM-hypointensities* only have 4 and 25 non-zero entries, which can indicate that they may not be useful for the models, but they were left in the dataset.

In Figure 4.1 we can see the correlation heatmap between all features. There are some emerging patterns which may be interesting to study in the future. By checking the features with correlation higher/lower than 0.95/-0.95 the following features were removed, which were assumed to be redundant, and with the correlation we confirmed that they do not contribute significant information: *CerebralWhiteMatterVol*, *BrainSegVolNotVent*, *BrainSegVolNotVentSurf*, *SupraTentorialVol* and *SupraTentorialVolNotVent*.

With this analysis nine features were removed from the dataset.

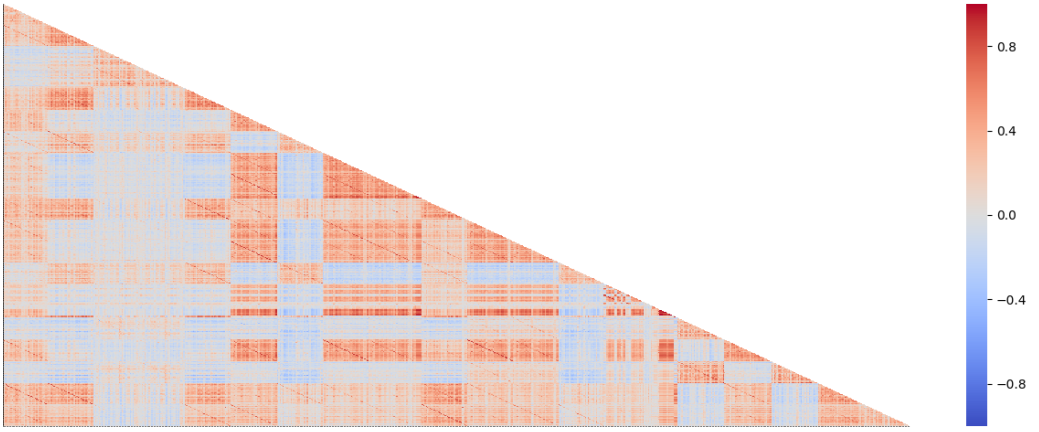


Figure 4.1: Correlation Heatmap.

4.3 Principal Component Analysis

4.3.1 PCA Results

In Figure 4.2 we can see the first 10 components, from applying PCA to the training test, and how using a combination of two of these components could be used to differentiate the label of DaTscan positive (blue) and negative (yellow). As we can see, there does not seem to be a pair of components that can be used to differentiate the label. Moreover, there seem to exist some subjects who could be considered outliers in regards to component 6.

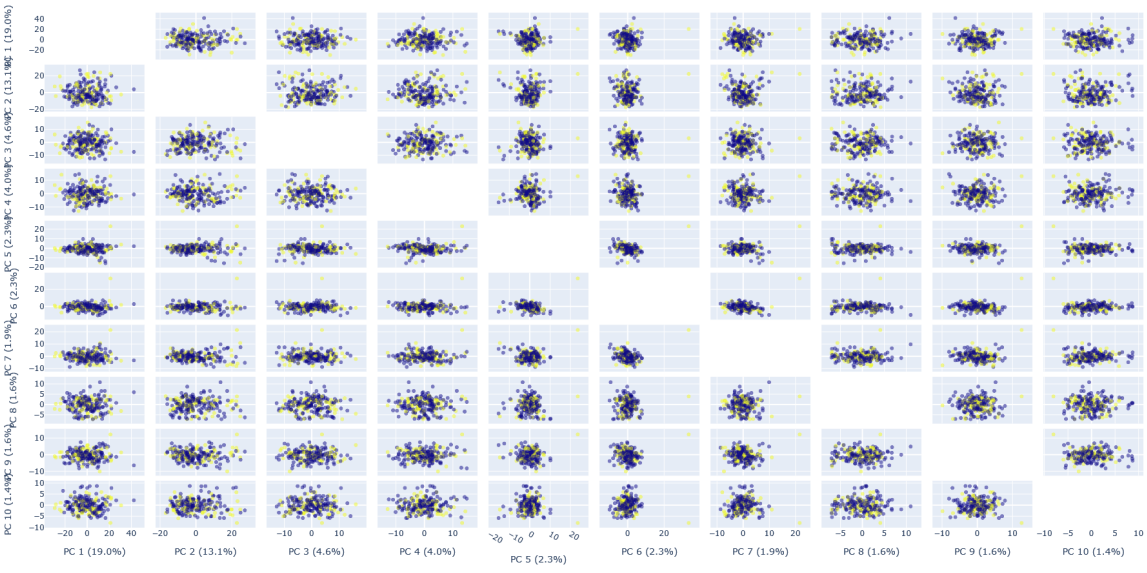


Figure 4.2: PCA first 10 components, with DaTscan label: positive (blue) and negative (yellow).

In terms of explained variance: 57, 95 and 216 features explain 80%, 90% and 100% respectively. The growth of the explained variance as a function of the number of features used can be seen in Figure 4.3.

Using the 57 components that explain 80% of variance, the initial features were assessed to see which ones have the highest contribution. The differences in explained variance between features can

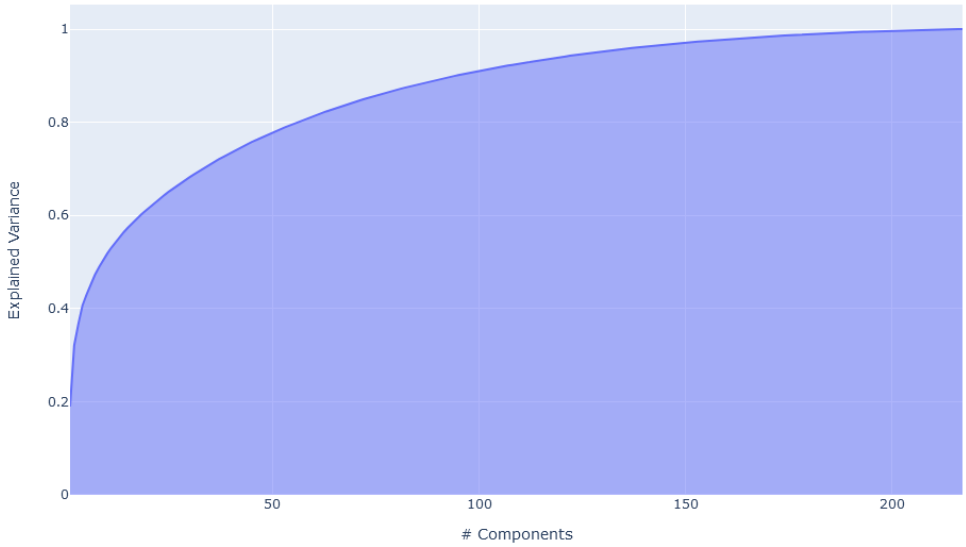


Figure 4.3: PCA explained variance.

be see in Figure 4.4, but since there are so many features, their names were omitted. The features with the highest contribution are the ones relative to the whole brain and not specific to any region, while most features that have low contribution are regions *thickness std* and the Ventricles features.

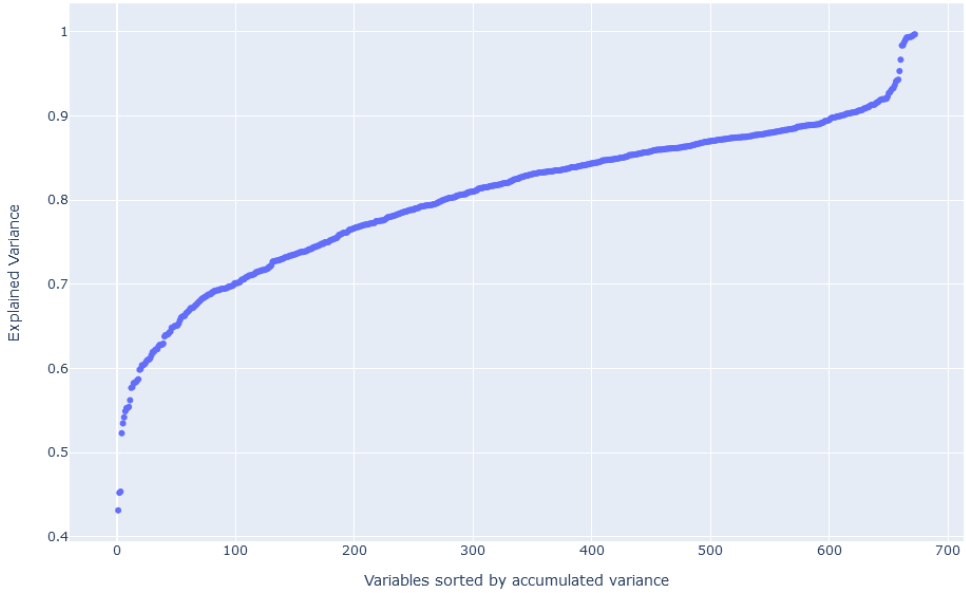


Figure 4.4: PCA features contribution for the first 57 components.

4.3.2 RPCA results

This section discusses similar results shown in section 4.3.1, but where RPCA was applied instead of PCA.

Figure 4.5 shows the first 10 components resulting from applying RPCA to the training test, and how using a combination of two of these components could be used to differentiate the label of DaTscan positive (blue) and negative (yellow). As we can see, there does not seem to be a pair of components

that can be used to differentiate the label. Moreover, in contrast to the PCA results, there do not seem the exist outliers affecting the results.

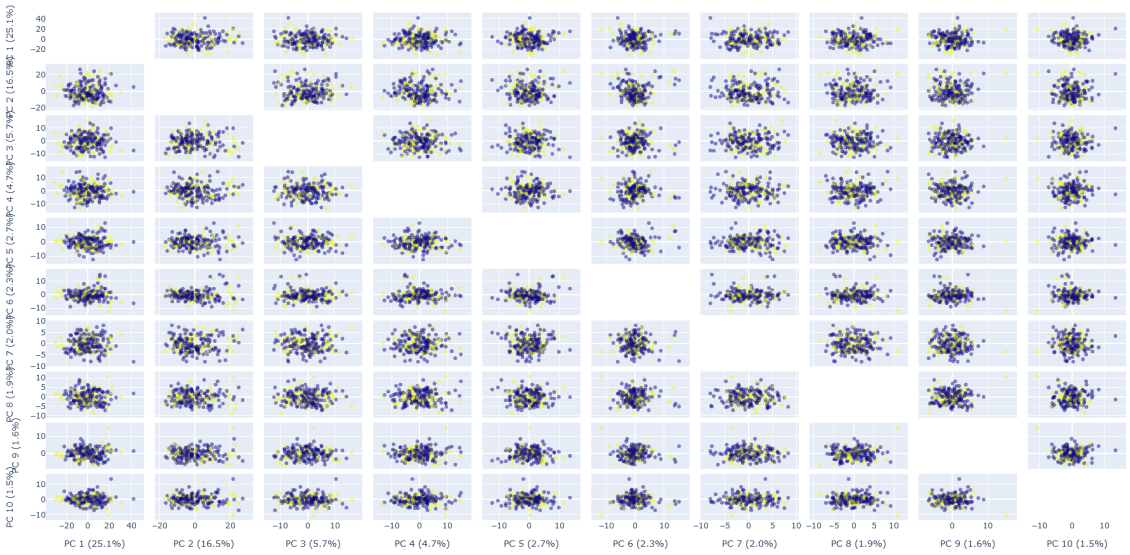


Figure 4.5: RPCA first 10 components, with DaTscan label: positive (blue) and negative (yellow).

In terms of explained variance: 29, 50 and 129 features explain 80%, 90% and 100% respectively. The graph for the evolution of the explained variance with more components can be seen in Figure 4.6.

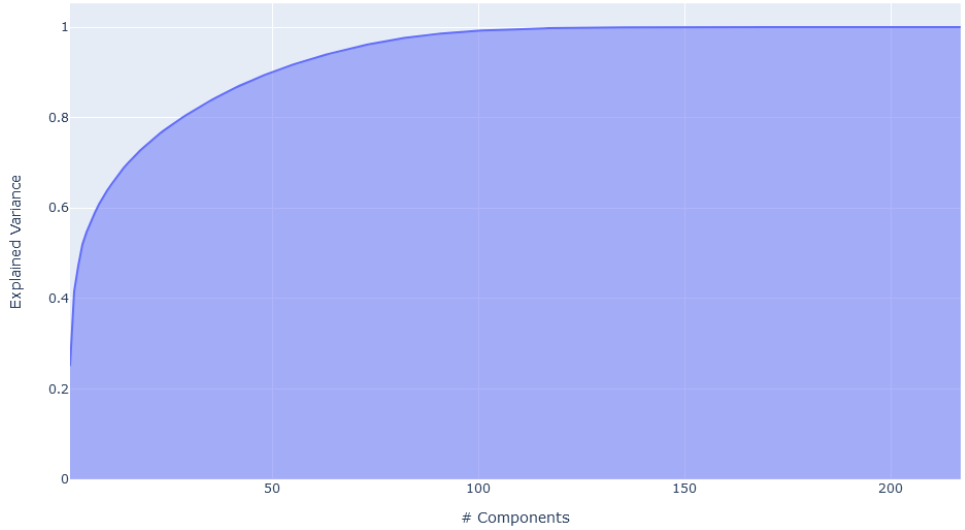


Figure 4.6: RPCA explained variance.

Using the 29 components that explain 80% of variance, the initial features were assessed to determine which are the ones that have the highest contribution. The differences in explained variance between features can be seen in Figure 4.7, but since there are so many features, their names are omitted. The features with highest and lowest contributions follow a similar pattern to the ones resulting from PCA, and in addition, the features relative to the regions *folding index* also have a low contribution.

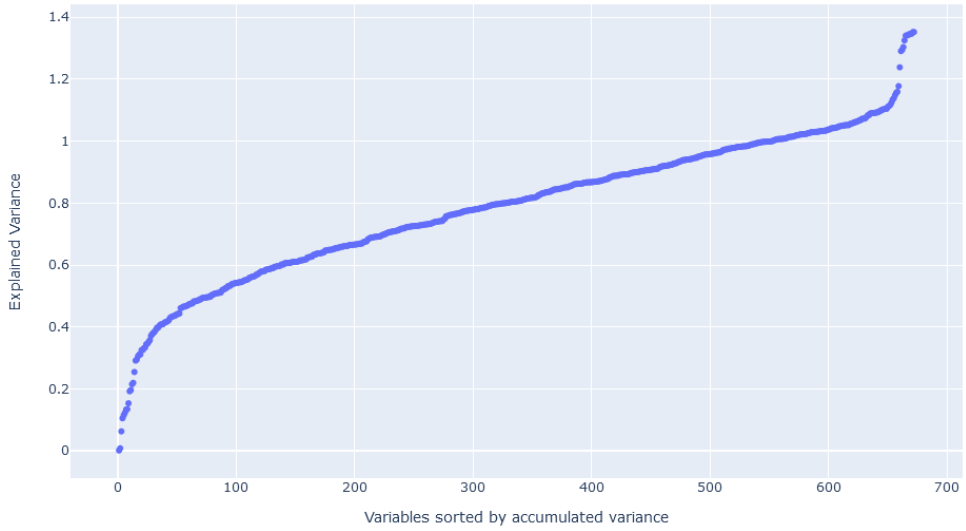


Figure 4.7: RPCA features contribution for the first 29 components.

4.4 Baseline Approach

Using the pipeline explained in Section 3.8 resulted in a model whose performance is shown in Table 4.6, with the results of labeling both the training dataset as well as the testing dataset. Across all metrics, we can see the model performed significantly worse in labelling the testing dataset when compared to the results from the training dataset. This is a clear sign that the model is overfitting the data, and a possible explanation would be the high complexity of the model and chosen hyperparameters.

Table 4.6: Results from baseline approach.

	Accuracy	**Balanced Accuracy**	**F1 score**	**Balanced F1 score**	**ROC AUC**
Train	96.6%	94.9%	93.3%	96.5%	99.5%
Test	54.5%	43.8%	16.7%	52.2%	40.0%

4.5 Exploration for Best Model Results

The results shown in this section are based on the following method: first we collect all the results and some statistics, then we compare details of the scores for the 5 best model combinations, for all features or regions and for validation or cross validation. Finally we review the details of the scores due to testing of these best models.

The balanced accuracy was used to compare all combinations, since using the accuracy could be misleading due to the data being unbalanced, but since accuracy is used in the literature this score was chosen instead of F1-score or ROC AUC.

4.5.1 General Results

In Figure 4.8 and 4.9 we see the legends for the symbols and colours used in Figure 4.10 and 4.11.

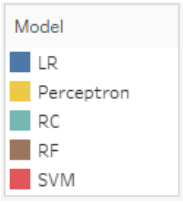


Figure 4.8: Model algorithm legend: Logistic regression (LR), Ridge Classifier (RC), Random Forest (RF), Support Vector Machine with rbf kernel (SVM)

Figure 4.9: Set/validation type legend

In Figure 4.10, we see symbols with colors that mark the balanced accuracy for all of the 160 combinations when using all features, depending on the choice of method detailed in Section 3.9. Each row corresponds to a different combination of balancing type and transformation. Note that multiple symbols may overlap.

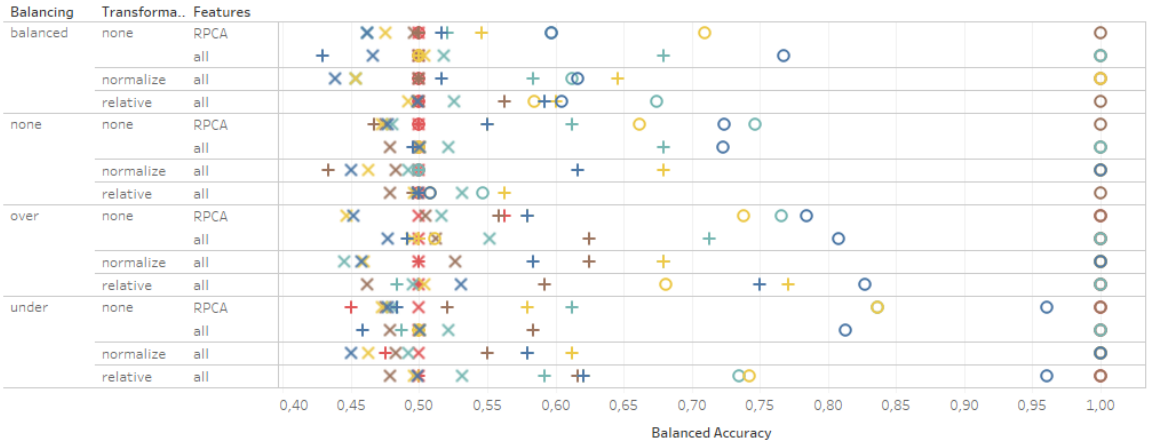


Figure 4.10: Results using all features

In Figure 4.11, we see symbols with colors that mark the balanced accuracy for all the 4920 combinations when using brain regions features subsets, with each row corresponding to one of these subsets.

There are more than 5000 possible combinations that are displayed as data points in these graphs, and it is not possible to discuss them individually. So, to be able to evaluate how different approaches affect the results, the balanced accuracy average was calculated for different aggregations, which are shown in Tables 4.7, 4.8, 4.9 with the combinations that used all the features, and Tables 4.10, 4.11 and 4.12 for the combinations that used subsets of features relative to brain regions.

Table 4.7: Average balanced accuracies for all features and for each algorithm

Model	Train	Validation	CV
LR	79.34%	54.77%	47.73%
Perceptron	71.70%	58.91%	48.09%
RC	81.33%	58.70%	50.13%
RF	100.00%	53.93%	48.97%
SVM	75.00%	49.92%	50.00%

Table 4.8: Average balanced accuracies for all features and for each balancing type

Balancing	Train	Validation	CV
None	72.09%	53.21%	49.01%
Undersampling	91.91%	54.60%	49.01%
Oversampling	90.58%	59.25%	49.18%
In-model	71.32%	53.92%	48.73%

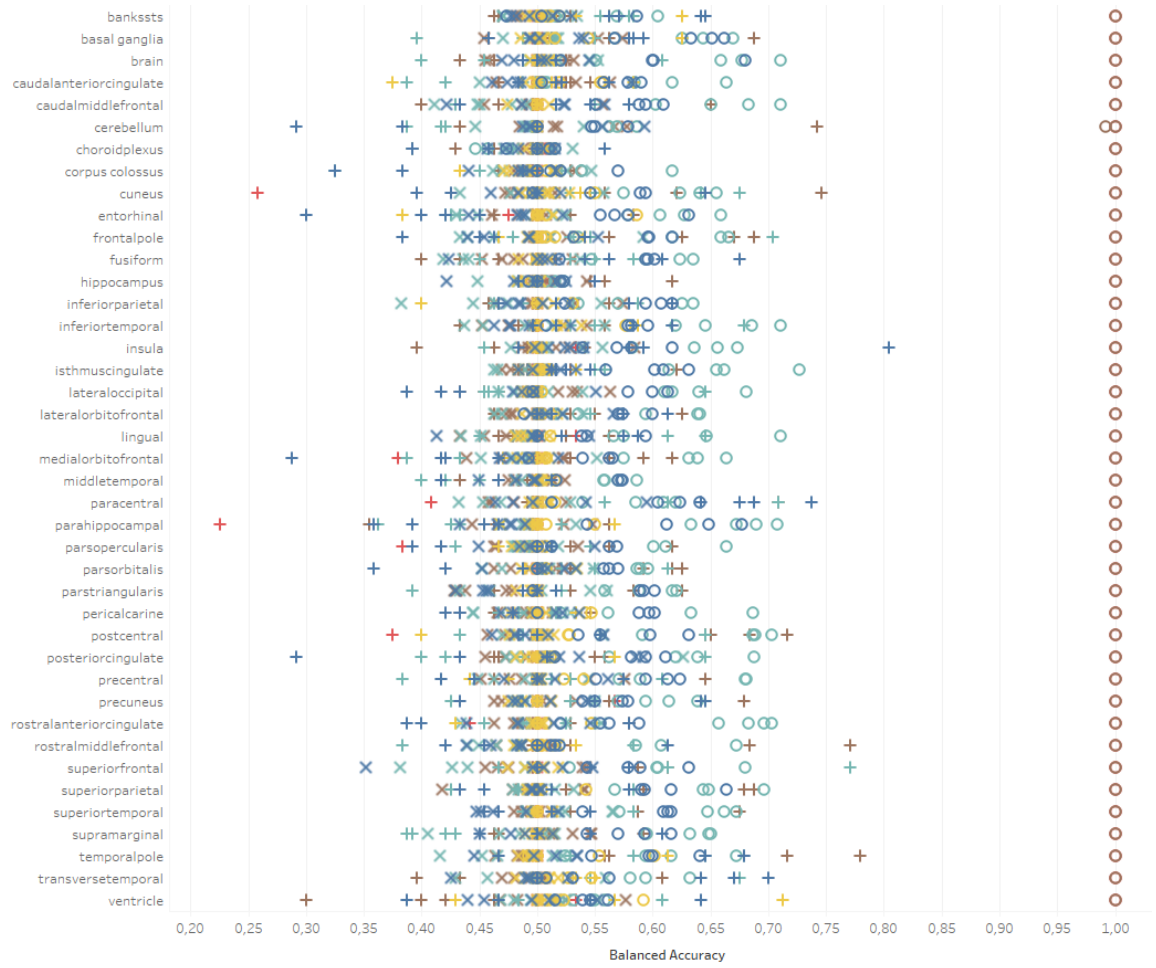


Figure 4.11: Results using subsets of features for brain region

Table 4.9: Average balanced accuracies for all features and for each transformation

Transformation	Train	Validation	CV
None	80.61%	53.21%	50.14%
RPCA	79.78%	53.83%	48.15%
Normalize	88.65%	56.69%	47.55%
Relative	76.86%	57.25%	50.10%

Table 4.10: Average balanced accuracies when using brain regions features subsets and for each algorithm

Model	Train	Validation	CV
LR	56.66%	50.67%	49.35%
Perceptron	50.83%	50.33%	49.88%
RC	60.76%	51.00%	49.05%
RF	100.00%	53.21%	49.47%
SVM	75.61%	49.47%	50.02%

4.5.2 Best Model Combinations Results

For the tables in this section, *B. Accuracy* represents balanced accuracy, and *B. F1-score* represents balanced F1-score.

In Table 4.13, we see the results for the 5 best combinations, (but since there were 4 equal, the 7

Table 4.11: Average balanced accuracies when using brain regions features subsets and for each balancing type

Balancing	Train	Validation	CV
None	61.96%	50.33%	49.70%
Undersampling	74.62%	50.80%	49.70%
Oversampling	74.27%	51.28%	49.97%
In-model	64.24%	51.33%	48.84%

Table 4.12: Average balanced accuracies when using brain regions features subsets and for each transformation

Transformation	Train	Validation	CV
None	68.77%	50.94%	49.64%
Normalize	68.77%	50.94%	49.50%
Relative	68.77%	50.94%	49.52%

best are shown) based on balanced accuracy, that used all features, and which were obtained using validation.

Table 4.13: Best combinations for all features, using validation

ID	Model	Transformation	Balancing	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC
1	Perceptron	relative	oversampling	77.08%	73.91%	74.52%	76.90%	77.08%
2	Logistic Regression	relative	oversampling	75.00%	82.61%	80.73%	88.20%	75.00%
3	Ridge Classifier	none	oversampling	71.25%	73.91%	73.91%	80.00%	71.25%
4	Ridge Classifier	none	in-model	67.92%	69.57%	69.94%	75.90%	67.92%
5	Ridge Classifier	none	none	67.92%	69.57%	69.94%	75.90%	67.92%
6	Perceptron	normalize	none	67.92%	69.57%	69.94%	75.90%	67.92%
7	Perceptron	normalize	oversampling	67.92%	69.57%	69.94%	75.90%	67.92%

In Table 4.14, we see the results for the 5 best combinations, based on balanced accuracy, that used all features, and which were obtained using cross-validation.

Table 4.14: Best combinations for all features, using cross-validation

ID	Model	Transformation	Balancing	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC
1	Ridge Classifier	none	oversampling	55.18%	52.92%	52.51%	56.10%	45.80%
2	Ridge Classifier	relative	none	53.15%	65.47%	57.50%	77.00%	53.15%
3	Ridge Classifier	relative	undersampling	53.15%	65.47%	57.50%	77.00%	53.15%
4	Logistic Regression	relative	oversampling	53.06%	45.97%	54.16%	57.30%	49.09%
5	Random Forest	normalize	oversampling	52.66%	59.35%	57.13%	77.20%	55.22%

In Table 4.15, we see the results for the 5 regions that gave the highest balanced accuracy, using validation. All the results were the same using the three possible transformations: none, relative and normalize.

Table 4.15: Best combinations, using brain regions features subsets and validation

Region	Model	Balancing	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC
insula	Logistic Regression	oversampling	80.42%	78.26%	78.77%	81.50%	80.42%
temporalpole	Random Forest	undersampling	77.92%	82.61%	81.91%	87.50%	77.92%
rostralmiddlefrontal	Random Forest	undersampling	77.08%	73.91%	74.52%	76.90%	77.08%
superiorfrontal	Ridge Classifier	in-model	77.08%	73.91%	74.52%	76.90%	77.08%
cuneus	Random Forest	oversampling	74.58%	78.26%	77.89%	83.90%	74.58%

In Table 4.16, we see the results for the 5 regions that gave the highest balanced accuracy, using cross-validation. The * in transformation indicates that the results were the same using the three possible transformations: none, relative and normalize, for that given region.

Table 4.16: Best combinations, using brain regions features subsets and with cross-validation

Region	Model	Transformation	Balancing	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC
posteriorcingulate	Ridge Classifier	*	in-model	62.67%	62.68%	63.68%	69.1%	62.67%
paracentral	Logistic Regression	normalize	oversampling	59.51%	58.14%	46.34%	59.5%	55.12%
cerebellum	Logistic Regression	normalize	oversampling	59.30%	51.15%	47.11%	58.00%	53.82%
posteriorcingulate	Logistic Regression	*	in-model	59.17%	59.01%	60.13%	65.40%	59.17%
insula	Random Forest	normalize	oversampling	58.43%	59.46%	58.41%	70.4%	48.52%

4.5.3 Testing the Best Model Combinations

The models from the previous section were tested, which produced the results shown in Tables 4.17, 4.18, 4.15 and 4.16. The True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN) values make up the confusion matrix. The Train B. Accuracy, is the score of using the entire train set to train these models with the choices obtained with either validation or cross-validation.

Table 4.17: Testing models from Table 4.13, that use all features, obtained with validation

ID	Train B. Accuracy	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC	TP	FN	FP	TN
1	61.34%	50.60%	62.77%	56.37%	75.86%	50.60%	55	7	28	4
2	85.86%	48.49%	50.00%	51.23%	58.41%	48.49%	33	29	18	14
3	100%	45.41%	48.94%	49.83%	45.41%	45.41%	35	27	21	11
4	100%	43.00%	45.74%	46.97%	55.65%	43.00%	32	30	21	11
5	100%	44.60%	47.87%	48.89%	58.12%	44.60%	34	28	21	11
6	100%	45.36%	47.87%	49.05%	57.39%	45.36%	33	29	20	12
7	98.97%	44.60%	47.87%	48.88%	58.12%	44.61%	34	28	21	11

Table 4.18: Testing models from Table 4.14, that use all features, obtained with cross-validation

ID	Train B. Accuracy	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC	TP	FN	FP	TN
1	100%	45.41%	48.94%	49.83%	59.32%	45.41%	35	27	21	11
2	52.08%	49.19%	64.89%	51.91%	78.71%	49.19%	61	1	32	0
3	100%	48.44%	48.94%	50.27%	56.36%	48.44%	31	31	17	15
4	85.86%	48.49%	50.00%	51.23%	58.41%	48.49%	33	29	18	14
5	100%	48.34%	62.77%	52.51%	76.82%	48.34%	58	4	31	1

Table 4.19: Testing models from Table 4.19, that use brain regions, obtained with validation

Region	Train B. Accuracy	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC	TP	FN	FP	TN
insula	56.90%	45.51%	51.06%	51.06%	62.90%	45.51%	39	23	23	9
temporalpole	100%	57.11%	56.38%	57.52%	62.39%	57.11%	34	28	13	19
rostralmiddlefrontal	100%	54.94%	58.51%	58.93%	67.77%	54.94%	41	21	18	14
superiorfrontal	57.45%	39.01%	41.49%	42.93%	51.33%	39.01%	29	33	22	10
cuneus	100%	47.33%	57.45%	53.65%	71.01%	47.33%	49	13	27	5

Table 4.20: Testing models from Table 4.20, that use brain regions, obtained with cross-validation

Region	Train B. Accuracy	B. Accuracy	Accuracy	B. F1-score	F1-score	ROC AUC	TP	FN	FP	TN
posteriorcingulate RC	64.36	45.26%	45.74%	47.16%	53.21%	45.26%	29	33	18	14
paracentral	53.10%	54.08%	56.38%	57.23%	64.96%	54.08%	38	24	17	15
cerebellum	58.28%	53.18%	53.19%	54.41%	60.00%	53.18%	33	29	15	17
posteriorcingulate LR	63.68%	46.77%	45.74%	47.05%	51.43%	46.77%	27	35	16	16
insula	100%	47.33%	57.45%	53.65%	71.01%	47.33%	49	13	27	5

4.6 Discussion

In regards to general results, a common trend across Tables 4.7 through 4.12 is that the balanced accuracy of the models is higher for the training set than for the validation set, with the latter being on

average slightly above 50%. This shows that most models suffered from overfitting. Tables 4.7 and 4.10 in particular show that models with algorithm Random Forest are especially prone to this, seeing as the average balanced accuracy for the training set is 100%.

Tables 4.8 and 4.11 indicate that the oversampling strategy for balancing data tends to give better results, both from the lens of validation and cross-validation. Tables 4.9 and 4.12 seem to indicate that there is no significant improvement from using any of the studied feature transformations.

Looking at the best models, Tables 4.13 and 4.14 suggest the Ridge Classifier algorithm provides the best models for dealing with a large number of features, whereas Tables 4.15 and 4.16 indicate Logistic Regression is most appropriate for the situations with fewer features. On the other hand, models using SVM as their algorithm are not represented in the top models. Regarding data balancing, oversampling seems to lead to better models, whereas RPCA is not represented among the top models. Finally, the features from the brain regions *insula* and *temporalpole* seem to be of importance for this classification.

Although these results may seem promising, it should be noted that, since we have studied so many combinations, finding strong results might simply be due to the likelihood of finding some model that happens to work well with this particular dataset. This is why it is important to use a testing dataset, which is completely separate from the pipeline followed up until this point.

The testing results described in Tables 4.17 through 4.20 show that, across all combinations, the top models selected before do not generalize well: for instance, the values of balanced accuracy do not go higher than 60%. Despite that, the majority of these models outclass the baseline considered for this work in terms of the balanced accuracy, as well as the other metrics calculated.

4.6.1 Validation vs Cross-Validation

Comparison of the validation and cross-validation results serves to further reinforce the findings of existing research. Across the top models tested, the balanced accuracy obtained from cross validation was a better predictor of the testing balanced accuracy, whereas the results of using the usual validation were overly optimistic. This is in line with observations suggesting cross validation should be preferred in situations where small datasets are available, as is the case here.

4.6.2 Limitations

The main conclusion is that, although we were able to obtain models that outperform the baseline, the approaches used with the given dataset are not able to separate between subjects with and without dopamine deficiency with the features that were extracted from a sMRI. This may have been caused by a few factors:

- The small size of the dataset may have caused the models not to have enough data to learn in a way that generalizes well. This is particularly true given the high complexity of the problem at hand, which is reflected in the high number of variables.

- The models chosen from Machine Learning may not have been adequate for this problem. Since

the models were chosen for their simplicity, the solution could be to apply more complex models. However, more complex models would require more datapoints to successfully separate the data.

- The two groups were not as homogeneous as it is seen in the literature. The group of SWEDD and Control were grouped as negative, to be able to view this problem as a binary one.
- The modality of the MRI and the features extracted may also affect what is possible to do with the data. Most articles published used DTI for problems similar to the hypothesis of this thesis.
- Finally, it may be the case that the task at hand cannot be performed with higher performance. There seem to be no other articles investigating this, so this is still an open question.

Chapter Summary

This chapter showed the results from analysis performed on the data and the results from the ML models used and tested. The results used to compare different approaches gave an indication that some distinction could be done of the dopamine deficiency and not deficiency, but when tested in the independent test group, these models showed no positive results.

Additionally, since the hypothesis was not able to be proven right, there was some discussion of why things might not have worked.

Chapter 5

Conclusions

The articles found and studied in this report show a lack of research done in using MRI to differentiate subjects with and without dopamine deficiency, and only a few using Diffusion MRI. However, the lab with which this thesis was produced has made some efforts in this area, using sMRI and deep learning. However, a lack of testing the simplest approach with the most common MRI, which is sMRI T1w, still remains. Consequently, the hypothesis of this thesis was to study whether it would be possible to use sMRI to identify which patients would have a positive result from a DaTscan, with the use of Machine Learning.

The accuracy score is used throughout literature to compare results, but with unbalanced data this can be very misleading. For instance, for this hypothesis we could achieve accuracies of 67% simply by reporting on a model that classified all as positive. For this reason balanced accuracy was chosen to be used as the score to compare approaches and report results.

With the initial analysis, some patterns in the data seem to emerge which could indicate that a smaller set of features could be used. Furthermore, the baseline ML approach seems to be overfitted, with balanced accuracy of the training and testing set being 94.6% and 43.8%, respectively. This indicated that a possible path would be to use simpler models and method of finding hyperparameters.

This thesis employed a variety of strategies to achieve and judge results with a validation set or cross-validation. These included: exploring different ML algorithms such as Logistic Regression, Perceptron, Ridge Classifier, Random Forest, and Support Vector Machine, different data transformations, such as RPCA, normalized data and transforming areas and volumes into relative features, alongside different balancing methods, such as oversampling, undersampling, and in model balancing of classes. Cross validation was a better predictor of the results achieved when further testing the models in the test set, when compared to the results judged with the validation set.

The different models, when further tested on an independent set, suffered from underfitting or overfitting, but with the best models having higher scores than the baseline approach. The highest result achieved was 50.60%.

The option of using subsets of features that were relative to brain regions all suffered from the same issues, yet this would be a method of feature selection that could pinpoint which regions would be of most

importance to this classification problem. The highest result achieved was 57.11% balanced accuracy with features from the *temporalpole*.

The main conclusion is that, although the study was able to obtain models that outperform the baseline, the approaches used with the given dataset are not able to differentiate between subjects with and without dopamine deficiency with the features that were extracted from a sMRI.

5.1 Future Work

Multiple paths can be followed to further test if it is possible to use MRI to identify which patients would have a positive result from a DaTscan. One of more important ones would be to gather more data, which is an effort being pursued by Diana Prata's Lab, although medical data and MRIs are not a category of data which is easily scalable. Using more homogeneous groups can be advantageous, by separating the negative group, and then solving three binary problems (PD vs SWEDD, Control vs SWEDD, PD vs Control) . However, this would make the data available even more unbalanced.

Another option is to consider using other types of features extracted from sMRI, or to use another modality such as DTI to try to replicate the results which exist in the literature. Furthermore, More sMRI can be used if when selecting them, less restrictions on the parameters are made. However, this this can present problems as a result of the images potentially being different enough that the models would detect them, and differentiate between them instead of the label that is intended.

Finally, more complex models, such as CNN or other deep learning methods, may bring more insights, but since there is not a lot of data, this is not a clear path to take.

Bibliography

[1] M. Robert A Hauser. Parkinson disease, Apr 2021. URL https://emedicine.medscape.com/article/1831191-overview#a2.

[2] C. Blauwendraat, M. A. Nalls, and A. B. Singleton. The genetic architecture of parkinson's disease. *The Lancet Neurology*, 19(2):170–178, Feb. 2020. doi: 10.1016/s1474-4422(19)30287-x. URL https://doi.org/10.1016/s1474-4422(19)30287-x.

[3] M. J. Armstrong and M. S. Okun. Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA*, 323(6):548–560, 02 2020. ISSN 0098-7484. doi: 10.1001/jama.2019.22360. URL https://doi.org/10.1001/jama.2019.22360.

[4] M. J. Farrer. Genetics of parkinson disease: paradigm shifts and future prospects. *Nature Reviews Genetics*, 7(4):306–318, Apr. 2006. doi: 10.1038/nrg1831. URL https://doi.org/10.1038/nrg1831.

[5] H. Braak, K. D. Tredici, U. Rüb, R. A. de Vos, E. N. Jansen Steur, and E. Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of Aging*, 24(2):197–211, 2003. ISSN 0197-4580. doi: https://doi.org/10.1016/S0197-4580(02)00065-9. URL https://www.sciencedirect.com/science/article/pii/S0197458002000659.

[6] L. V. Kalia and A. E. Lang. Parkinson's disease. *The Lancet*, 386(9996):896–912, Aug. 2015. doi: 10.1016/s0140-6736(14)61393-3. URL https://doi.org/10.1016/s0140-6736(14)61393-3.

[7] C. A. Haaxma, B. R. Bloem, G. F. Borm, W. J. G. Oyen, K. L. Leenders, S. Eshuis, J. Booij, D. E. Dluzen, and M. W. I. M. Horstink. Gender differences in parkinson's disease. 78(8):819–824, Aug. 2007. doi: 10.1136/jnnp.2006.103788. URL https://doi.org/10.1136/jnnp.2006.103788.

[8] J. Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376, 2008. ISSN 0022-3050. doi: 10.1136/jnnp.2007.131045. URL https://jnnp.bmj.com/content/79/4/368.

[9] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino. Accuracy of clinical diagnosis of parkinson disease. *Neurology*, 86(6):566–576, 2016. ISSN 0028-3878. doi: 10.1212/WNL.0000000000002350. URL https://n.neurology.org/content/86/6/566.

[10] R. de la Fuente-Fernandez. Role of DaTSCAN and clinical diagnosis in parkinson disease. *Neurology*, 78(10):696–701, Feb. 2012. doi: 10.1212/wnl.0b013e318248e520. URL https://doi.org/10.1212/wnl.0b013e318248e520.

[11] R. Erro, S. A. Schneider, M. Stamelou, N. P. Quinn, and K. P. Bhatia. What do patients with scans without evidence of dopaminergic deficit (SWEDD) have? new evidence and continuing controversies. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(3):319–323, May 2015. doi: 10.1136/jnnp-2014-310256. URL https://doi.org/10.1136/jnnp-2014-310256.

[12] Magnetic resonance imaging, Apr 2021. URL https://en.wikipedia.org/wiki/Magnetic_resonance_imaging.

[13] L. Chougar, N. Pyatigorskaya, B. Degos, D. Grabli, and S. Lehéricy. The role of magnetic resonance imaging for the diagnosis of atypical parkinsonism. *Frontiers in Neurology*, 11, July 2020. doi: 10.3389/fneur.2020.00665. URL https://doi.org/10.3389/fneur.2020.00665.

[14] Fsl course preparatory lecture - part 1. URL https://youtu.be/Y6Mu_09ou5E.

[15] R. Deb, G. Bhat, S. An, H. Shill, and U. Y. Ogras. Trends in technology usage for parkinson's disease assessment: A systematic review. Feb. 2021. doi: 10.1101/2021.02.01.21250939. URL https://doi.org/10.1101/2021.02.01.21250939.

[16] J. M. Valverde, V. Imani, A. Abdollahzadeh, R. D. Feo, M. Prakash, R. Ciszek, and J. Tohka. Transfer learning in magnetic resonance brain imaging: A systematic review. *Journal of Imaging*, 7(4):66, Apr. 2021. doi: 10.3390/jimaging7040066. URL https://doi.org/10.3390/jimaging7040066.

[17] A. Segato, A. Marzullo, F. Calimeri, and E. D. Momi. Artificial intelligence for brain diseases: A systematic review. *APL Bioengineering*, 4(4):041503, Dec. 2020. doi: 10.1063/5.0011697. URL https://doi.org/10.1063/5.0011697.

[18] J. M. Mateos-Pérez, M. Dadar, M. Lacalle-Aurioles, Y. Iturria-Medina, Y. Zeighami, and A. C. Evans. Structural neuroimaging as clinical predictor: A review of machine learning applications. *NeuroImage: Clinical*, 20:506–522, 2018. doi: 10.1016/j.nicl.2018.08.019. URL https://doi.org/10.1016/j.nicl.2018.08.019.

[19] J. Xu and M. Zhang. Use of magnetic resonance imaging and artificial intelligence in studies of diagnosis of parkinson's disease. *ACS Chemical Neuroscience*, 10(6):2658–2667, May 2019. doi: 10.1021/acschemneuro.9b00207. URL https://doi.org/10.1021/acschemneuro.9b00207.

[20] L. Zhang, M. Wang, M. Liu, and D. Zhang. A survey on deep learning for neuroimaging-based brain disorder analysis. *Frontiers in Neuroscience*, 14, Oct. 2020. doi: 10.3389/fnins.2020.00779. URL https://doi.org/10.3389/fnins.2020.00779.

[21] A. A.-A. Valliani, D. Ranti, and E. K. Oermann. Deep learning and neurology: A systematic review. *Neurology and Therapy*, 8(2):351–365, Aug. 2019. doi: 10.1007/s40120-019-00153-8. URL https://doi.org/10.1007/s40120-019-00153-8.

[22] A. D. Yao, D. L. Cheng, I. Pan, and F. Kitamura. Deep learning in neuroradiology: A systematic review of current algorithms and approaches for the new wave of imaging technology. *Radiology: Artificial Intelligence*, 2(2):e190026, Mar. 2020. doi: 10.1148/ryai.2020190026. URL https://doi.org/10.1148/ryai.2020190026.

[23] K. R. Bhatele and S. S. Bhadauria. Brain structural disorders detection and classification approaches: a review. *Artificial Intelligence Review*, 53(5):3349–3401, Oct. 2019. doi: 10.1007/s10462-019-09766-9. URL https://doi.org/10.1007/s10462-019-09766-9.

[24] E. U. Haq, J. Huang, L. Kang, H. U. Haq, and T. Zhan. Image-based state-of-the-art techniques for the identification and classification of brain diseases: a review. *Medical & Biological Engineering & Computing*, 58(11):2603–2620, Sept. 2020. doi: 10.1007/s11517-020-02256-z. URL https://doi.org/10.1007/s11517-020-02256-z.

[25] L. Jin, Q. Zeng, J. He, Y. Feng, S. Zhou, and Y. Wu. A ReliefF-SVM-based method for marking dopamine-based disease characteristics: A study on SWEDD and parkinson's disease. *Behavioural Brain Research*, 356:400–407, Jan. 2019. doi: 10.1016/j.bbr.2018.09.003. URL https://doi.org/10.1016/j.bbr.2018.09.003.

[26] E. Matsusue, Y. Fujihara, K. Tanaka, Y. Aozasa, M. Shimoda, H. Nakayasu, K. Nakamura, and T. Ogawa. The utility of the combined use of 123i-FP-CIT SPECT and neuromelanin MRI in differentiating parkinson's disease from other parkinsonian syndromes. *Acta Radiologica*, 60 (2):230–238, May 2018. doi: 10.1177/0284185118778871. URL https://doi.org/10.1177/0284185118778871.

[27] M. Kim and H. Park. Structural connectivity profile of scans without evidence of dopaminergic deficit (SWEDD) patients compared to normal controls and parkinson's disease patients. *SpringerPlus*, 5(1), Aug. 2016. doi: 10.1186/s40064-016-3110-8. URL https://doi.org/10.1186/s40064-016-3110-8.

[28] M. Kim and H. Park. Using tractography to distinguish SWEDD from parkinson's disease patients based on connectivity. *Parkinson's Disease*, 2016:1–10, 2016. doi: 10.1155/2016/8704910. URL https://doi.org/10.1155/2016/8704910.

[29] H. R. Pereira. *Classification of patients with parkinsonian syndromes using medical imaging and artificial intelligence algorithms*. Universidade Nova de Lisboa, 2018. URL https://run.unl.pt/handle/10362/61556.

[30] Principal component analysis (pca). URL https://scikit-learn.org/stable/modules/decomposition.html#principal-component-analysis-pca.

[31] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis?, 2009.

[32] Robust pca. URL https://github.com/amueller/advanced_training/blob/master/robust_pca.py.

[33] Random forests. URL https://scikit-learn.org/stable/modules/ensemble.html#forest.

[34] Support vector machines. URL https://scikit-learn.org/stable/modules/svm.html.

[35] Logistic regression. URL https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

[36] Perceptron. URL https://scikit-learn.org/stable/modules/linear_model.html#perceptron.

[37] Ridge regression and classification. URL https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression.

[38] Ppmi. URL https://www.ppmi-info.org/.

[39] Freesurfer, Apr 2021. URL https://surfer.nmr.mgh.harvard.edu/fswiki.

[40] Freesurfer pipeline, Apr 2021. URL https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferAnalysisPipelineOverview.

[41] V. Tavares, D. Prata, and H. A. Ferreira. Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and alzheimer's disease study. *Journal of Neuroscience Methods*, 334:108565, Mar. 2020. doi: 10.1016/j.jneumeth.2019.108565. URL https://doi.org/10.1016/j.jneumeth.2019.108565.

[42] C. A. Haaxma, B. R. Bloem, G. F. Borm, W. J. G. Oyen, K. L. Leenders, S. Eshuis, J. Booij, D. E. Dluzen, and M. W. I. M. Horstink. Gender differences in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(8):819–824, Aug. 2007. doi: 10.1136/jnnp.2006.103788. URL https://doi.org/10.1136/jnnp.2006.103788.