2018-09-19

Automatic Classification of Idiopathic Parkinsonian Disease and Progressive Supranuclear Palsy using Multi-Spectral MRI Datasets: A Machine Learning Approach

Talai, Aron Sahand

UNIVERSITY OF CALGARY

Automatic Classification of Idiopathic Parkinsonian Disease and Progressive Supranuclear Palsy using Multi-Spectral MRI Datasets: A Machine Learning Approach

by

Aron Sahand Talai

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN BIOMEDICAL ENGINEERING

CALGARY, ALBERTA

SEPTEMBER, 2018

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Automatic classification of idiopathic Parkinsonian disease and progressive supranuclear palsy using multi-spectral MRI datasets: A machine learning approach" submitted by Aron Sahand Talai in partial fulfillment of the requirements of the degree of MASTER OF SCIENCE.

ABSTRACT

Parkinson's disease, which is characterized by a range of motor and non-motor symptoms is categorized into classical Parkinsonian disease (PD) and atypical Parkinsonian syndromes (APS), such as progressive supranuclear palsy Richardson's syndrome (PSP-RS). The differential diagnosis between PD and PSP-RS is often challenged by similarity of early symptoms, effectively resulting in considerable misclassification rates. The aim of this thesis is to assess the benefits of using biomarkers from multi-modal MRI datasets in the accurate classification of PD vs. PSP-RS.

Multi-spectral information form T1-, T2-, and diffusion-weighted (DWI) MRI from 38 healthy controls (HC), 45 PD, and 20 PSP-RS subjects were available for this study. In detail, morphological (category 1), brain iron marker (category 2), and diffusion features (category 3) were employed. In the last category, all feature types were combined (combinational) for the development of a machine learning model. Nested leave-one-out-cross validation was used to evaluate the classification performance in each category followed by a 1000 permutation test to assess classification significance. The results suggest that, the DWI based classifier tied with the combinational approach in terms of overall accuracy. However, in the former, the specificity was lower by 10%. In detail, 4 PSP-RS and 1 PD subjects are incorrectly classified as PD and PSP-RS in the combinational approach resulting in a sensitivity and specificity of 91.67% and 94.12%, respectively. The obtained results indicate that features extracted from T1- and T2-weighted MRI perform worst based on overall accuracy. All classification categories were statistically significant ($p<0.001$).

In conclusion, combination of features from different MRI modalities such as T1-, T2-, and diffusion-weighted datasets improves the multi-level classification performance of HC vs. PD vs.

PSP-RS compared to single modality features, particularly in terms of PD vs. other differentiation. The results and concepts discussed in this research thesis have wide ranging implication for future developments of computer-aided diagnosis of PD sub-syndromes.

PREFACE

This thesis is original, unpublished, independent work by the author Aron S. Talai.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Nils Forkert, whose expertise, teaching style, financial and emotional support added immensely to my overall graduate experience. I appreciate his knowledge, technical skills, and above all, his mentor-ship style which has pushed me to go beyond what is expected from a typical M.Sc. student. Surely, in this highly competitive day and age, one must push their boundaries as much as possible as it will arm you with a better trained and progressed mind. I would like to thank the other members of my committee, Drs. Monchi and Chan for the oversight they provided at all levels of this research project. Ultimately, I would also like to thank my parents, both highly decorated academics in their respective fields, for their unconditional support throughout my life. I will never be able to convey my appreciation fully as I owe everything I have to them.

DEDICATION

To my lovely parents, Ziba and Hasan.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

The classical variant of Parkinson's disease (PD) and progressive supranuclear palsy (PSP) are characterized by progressive cell loss in the form of atrophy and other degenerative processes. While the underlying physiological aspects of these two entities are widely different, the manifestation profile of early stage symptoms are similar, making a correct differentiation a challenging task. Current standard clinical differentiation methods of PD and PSP are based on clinical questioners, however, due to the similarity of symptoms, such methods are not conclusive enough, as 25% of misclassification have been reported in non-specialized movement disorder clinics[1]. However, it is worth noting that while the differentiation between PD and PSP is challenging in the early stages, these entities present more unique symptoms as they progress. In this context, accurate clinical differentiation at an early stage enables proper administration of medicine that could potentially improve prognosis.

In order to reduce such misclassification rates, current studies have been focused on employing medical imaging approaches such as magnetic resonance imaging (MRI). Consequently, numerous group-wise and individual level studies have been investigated in the past. In terms of group-wise studies, previous research has been focused on identifying differences between PD and PSP using morphological, brain iron markers, and tissue integrity measurements via T1-weighted MRI, T2-weighted MRI, and diffusion-weighted MRI (DWI), respectively. Moreover, in terms of individual level studies, researchers have utilized the aforementioned measurements in the context of machine learning methods to perform binary or multi-level classifications of healthy controls (HC), PD, and PSP.

In short, machine learning approaches in the context of PD vs. PSP, take advantage of a set of datasets (e.g. T1-weighted images of individuals with PD and PSP) that have been pre-labeled by movement disorder specialists, with a set number of identifying features (i.e. morphological measurements that can derived from T1-weighted MRI). Machine learning methods aim to train a mathematical model based on the identifying features to classify future unseen data instances, thus providing an individual level classification. Most of the individual level studies in PD vs. PSP have focused on employing single channel MRI features such as morphology or brain iron marker measurements. However, only a few of these studies have utilized information form one, two or more of these imaging methods. Previous research in similar fields have indicated potential classification performance benefits from combining features from multiple modalities as they provide a more comprehensive profile of the diseases entities compared to single modality features.

1.1 Thesis Organization

In this section, a brief overview of the chapters in this research thesis will be provided. In chapter two: background, the fundamental aspects of this research project are explained in detail. In this chapter, the origins of the parkinsonian syndrome and current differential diagnosis methods of PD and PSP will be discussed. Moreover, an in-depth explanation of various MRI modalities used in this study as well as current state of the art group-wise and individual level studies in Parkinson's disease will be discussed. In this context, common processing methods often employed in neuroimaging such as registration and feature extractions will be elaborated on. Furthermore, a comprehensive overview of machine learning algorithms and related techniques will be discussed.

In chapter three: objectives and hypothesis, the goals and objectives of this research will be clearly defined. Furthermore, the potential implications of this research in the broader context of Parkinson's disease diagnosis advancements will be outlined.

In chapter four: materials and methods, the study cohort demographics as well as MRI imaging parameters are thoroughly explained. Moreover, in this chapter, the various pre-processing approaches that were used to accurately normalize medical images as well as the extraction of the required features from the raw MRI image sequence. Most importantly, this chapter will detail the core aspect of this project, which is the machine learning pipeline and some of the rational with regards to the implementation of this method. In detail, the pipeline consists of a feature selection block and a so-called classification block. Overall, the feature selection block contains numerous feature selection algorithms, whereas the classification block implements various classification algorithms such as random forest, support vector machines and others.

In chapter five: results and analysis, the results of the analysis using the aforementioned machine learning pipeline with respect to the differentiation of PD and PSP will be described.

Ultimately in the sixth and last chapter: discussion and conclusions, the results of this study will be rationalized by comparing them with previous research in the field. An in-depth critical analysis of the results will be conducted to showcase the benefits of using a combination of features from multi-modal MRI modalities. Moreover, this chapter will also contain a detailed category-specific analysis that highlights the key differences between these modalities and how they can be leveraged for accurate healthy control vs. PD vs. PSP classifications. Furthermore, the potential future directions as well as the limitations of this study will be discussed. This

chapter and thesis is then concluded by a comprehensive conclusion highlighting the achievements of this project and how these can be utilized for future studies.

CHAPTER TWO: BACKGROUND

2.1 Pathophysiology of the Parkinsonian Syndrome

The classical variant of Parkinson's disease (PD) is one of the most prevalent neurodegenerative movement disorders.[1] The primary cause of this chronic-progressive disease is typically accredited to accumulation of alpha-synuclein and progressive loss of dopaminergic cells within the substantia nigra.[2] In detail, progressive neuronal loss in the pars compacta of the substantia nigra and also in other pigmented nuclei[3] followed by white matter vitiation has been connected to PD.[4,5] The mechanism of cell loss in Parkinson's disease is thought to be caused by mitochondrial, oxidative stress, and abnormal handling of protein. These alterations are primarily perpetuated by age, which is an important risk factor in PD.[6]

PD is clinically characterized by a broad range of motor symptoms, including bradykinesia, asymmetric rigidity, resting tremor, and postural instability, as well as non-motor symptoms including hyposmia, mental and cognitive impairments, depression, constipation, or rapid eye movement sleep behavior disorders.[7] Levodopa and dopamine agonists are the main medical interventions for controlling motor related symptoms, whereas anti-depression medication and clozapine are said to mitigate depression and hallucinations symptoms, respectively.[7] Life expectancy in PD is similar to individuals without PD, however, patients are prone to falls, pneumonia, and other life debilitating incidents that drastically reduce the quality of life. Overall as of now, there are no clinical procedures to stop PD progression. However, some common therapies to impede disease advancement or mitigate symptoms include cell transplantation and deep brain stimulation.[8]

In terms of disease prevalence, an estimated 7 to 10 million people worldwide are living with Parkinson's disease. Fifty-five thousand Canadians aged 18 or older are reported to have Parkinson's disease, whereas 97% of them are older than 65.[9] Overall, statistical data suggests that men are 1.5 times more likely to develop PD than women.[9] The combined direct and indirect costs of PD in the US are reported to be up to 3000$ per year.[10]

In contrast, progressive supranuclear palsy Richardson's syndrome (PSP-RS), an atypical variant of the Parkinsonian syndrome, which belongs histo-pathologically to the tauopathies, is characterized by vertical supranuclear gaze palsy or slow velocity of vertical saccades, axial rigidity, and repeated unprovoked falls in the early disease course. From an anatomical perspective, the basal ganglia, diencephalon, brainstem, and the thalamus are impacted in PSP-RS.[11] In detail, the intra-cerebral aggregation of tau proteins is considered to be due to the H1-haplotype of the so-called MAPT gene.[12,13] Impacted neurons in this sporadic multi-system degenerative disease have a dense filamentous aggregates profile similar to the globose neurofibrillary tangles.[14]

According to previous studies, PSP's incidence is 30 times less than PD. However, with a mean onset age of 63 and a life expectancy of only up to 9 years, PSP-RS has a much less optimistic prognosis than PD.[15] As of now, no clinically proven medication is available for the treatment of PSP-RS. Several interventions such as davunetide, glycogen synthase kinase-3, lithium, riluzole, and lisuride have been investigated in the past with no significant success.[15] The lack of pharmacological options has shifted disease management onto relieving symptoms as well as improving the patient's quality of life.

2.1.1 Clinical Differentiation of Parkinsonian Syndromes

Differential diagnosis of Parkinson's disease is commonly conducted by symptom monitoring, with clinical assessments based on criteria defined by the UK Parkinson's Disease Society Brain Bank,[16] the US National Institute of Neurological Disorders and stroke,[17] and the Unified Parkinson's Disease Rating Scale (UPDRS)[18] in combination with magnetic resonance brain imaging.[19] For example, scores related to cognitive impairments, sleep problems as well as motor related issue such as walking and hand movements can be recorded by the patient and the health care professional. Figure 1.1 depicts a segment of the UPDRS scoring system.

MDS UPDRS Score Sheet

1.A	Source of information	☐ Patient ☐ Caregiver ☐ Patient + Caregiver	3.3b	Rigidity– RUE	
			3.3c	Rigidity– LUE	
Part I			3.3d	Rigidity– RLE	
1.1	Cognitive impairment		3.3e	Rigidity– LLE	
1.2	Hallucinations and psychosis		3.4a	Finger tapping– Right hand	
1.3	Depressed mood		3.4b	Finger tapping– Left hand	
1.4	Anxious mood		3.5a	Hand movements– Right hand	
1.5	Apathy		3.5b	Hand movements– Left hand	
1.6	Features of DDS		3.6a	Pronation- supination movements– Right hand	
1.6a	Who is filling out questionnaire	☐ Patient ☐ Caregiver ☐ Patient + Caregiver	3.6b	Pronation- supination movements– Left hand	
			3.7a	Toe tapping–Right foot	
1.7	Sleep problems		3.7b	Toe tapping– Left foot	
1.8	Daytime sleepiness		3.8a	Leg agility– Right leg	
1.9	Pain and other sensations		3.8b	Leg agility– Left leg	
1.10	Urinary problems		3.9	Arising from chair	
1.11	Constipation problems		3.10	Gait	
1.12	Light headedness on standing		3.11	Freezing of gait	
1.13	Fatigue		3.12	Postural stability	
Part II			3.13	Posture	
2.1	Speech		3.14	Global spontaneity of movement	
2.2	Saliva and drooling		3.15a	Postural tremor– Right hand	
2.3	Chewing and swallowing		3.15b	Postural tremor– Left hand	
2.4	Eating tasks		3.16a	Kinetic tremor– Right hand	
2.5	Dressing		3.16b	Kinetic tremor– Left hand	
2.6	Hygiene		3.17a	Rest tremor amplitude– RUE	

FIGURE 1.1. A SNAPSHOT OF THE UNIFIED PARKINSON'S DISEASE RATING SCALE (UPDRS) DIAGNOSIS CRITERIA

However, the main drawback of these diagnosis methods is that syndrome-specific symptoms between PD and the various atypical Parkinsonian subtypes overlap, especially in the early

disease stages.[20] This might cause inaccurate initial diagnosis, which in turn, causes suboptimal disease prognosis. In terms of diagnosis through magnetic resonance imaging, PD and PSP-RS often show noticeable structural differences in several regions such as the midbrain (also known as the hummingbird or the penguin sign), superior cerebellar peduncles, and the dilatation of the 4th ventricle (also known as the Mickey Mouse sign).[21] However, these structural alterations are not necessarily obvious in the early disease stages and they might also be subject to visual misinterpretation by the clinician. Figure 1.2 depicts the penguin and Mickey Mouse signs, which are often regarded as the clinical hallmark of PSP-RS.

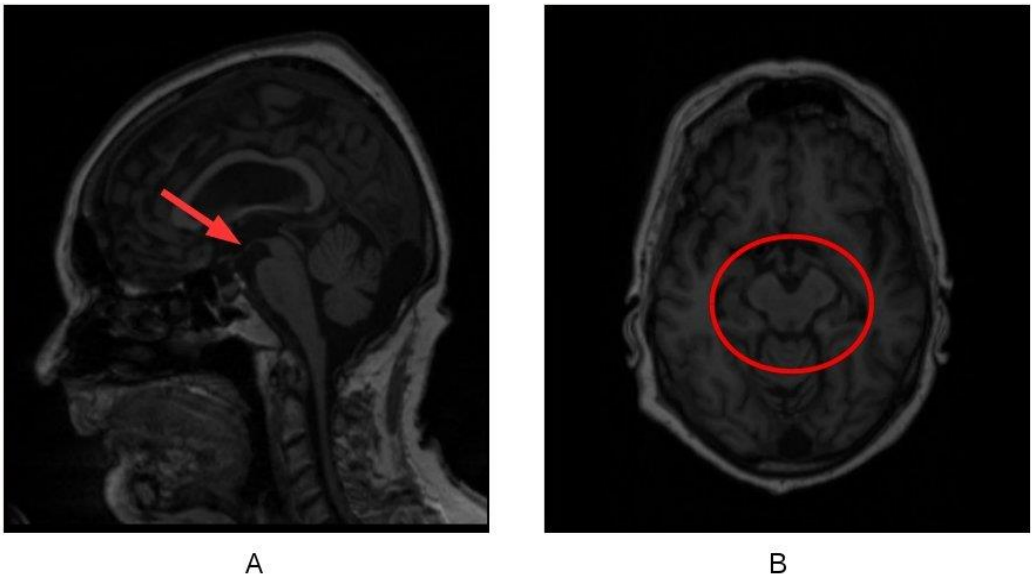

FIGURE 1.2. STRUCTURAL REPRESENTATIONS OF PSP-RS USED FOR DIAGNOSIS OFTEN IN CONJUNCTION WITH CLINICAL CRITERIA (A) THE PENGUIN SIGN AND (B) THE MICKEY MOUSE SIGN.

The correct recognition of PD is of the utmost importance since early stage syndrome-specific medicinal intervention has been shown to mitigate symptoms significantly.[22] Unfortunately, up to 25% of initial diagnoses of PSP are falsely declared as PD, which is a significant margin of

error.[23] It should be noted that highly specialized movement disorder centers are usually capable of differentiating different PD syndromes with considerably higher accuracy. However, not all patients have access to those specialized centers and even those centers do not achieve a 100% diagnostic accuracy. As previously mentioned, while there are currently no medical interventions targeting PSP-RS, an early differential diagnosis scheme will certainly improve disease prognosis as novel medical interventions are developed in the future.

2.2 MRI Basics

In this section, a brief overview of the fundamentals of the magnetic resonance imaging modality is provided. In-depth description of the specific MRI protocols employed in this study and their significance with regards to the Parkinsonian syndrome are discussed. As the focus of this thesis is not MRI sequence development, only core principles are described. Similar to currently existing imaging modalities, MRI image formation is fundamentally based upon creating contrast between various tissue types within a certain region of interest. In fact, the reason MRI (specifically T1-wighted MRI) is often employed in neurological brain disorders is the superior soft tissue detail it provides compared to other widely available imaging modalities such as X-ray computed tomography. MRI's flexibility and relative safety compared to other modalities has solidified its standing as one of the most employed imaging modalities in brain related studies such as Parkinson's disease.

2.2.1 MRI basics

Magnetic resonance imaging produces images (contrast) by the manipulation of nuclear angular momentum, spin frequencies, and other factors related to hydrogen atoms. The reason MRI is

mainly focused on the hydrogen atom is its single proton profile as well as it's abundance in human and animals in the form of water molecules. A certain region of the body (e.g. brain) will contain “N” number of hydrogen atoms, which is denoted by a factor called proton density. Each of these protons will have a certain local magnetic field with random directions as well as a constant movement which resembles the movement of a “top” called precessesion. It turns out that the overall combined magnetic influence of these N protons is zero. All of the above assumptions hold true in the absence of any magnetic interference, which is also called the equilibrium state. Now, if an external magnetic field (e.g. the main magnet of the MRI machine, producing a field strength of B0) is activated, the atom's individual magnetization vectors as well as spin profile will align, either parallel or anti parallel, with the external field. A slight majority of the atoms will be parallel to the B0 field, whereas the rest will be anti-parallel, effectively changing the net magnetization profile to a non-zero vector value M0.

In order to further excite hydrogen atoms and measure any profile changes, one can apply another magnetic field (radio frequency field) perpendicular to the B0 field, here called the B1 field. These fields are typically generated via the so-called gradient coils. In this context, the B1 magnetic field needs to have the same Larmor frequency as the aforementioned hydrogen atoms under the influence of the B0 field to initialize excitation. The Larmor frequency is defined in equation 1.1 where y is the so-called gyromagnetic ratio specific to the type of atom under investigation (y for Hydrogen equals 42.58 MHz/T).

$$w = y \times B \quad (1.1)$$

In detail, the B1 field is created via the application of a radio frequency (RF) pulse, which changes the M0 net vector based on its application angle, strength, duration, and other factors. For instance, a typical set-up would be an application angle (also called a flip angle) of 90 degrees with micro Tesla strength levels administered in a few milliseconds. After the B1 field is applied, the altered individual magnetization vector of the protons precesses at the Larmor frequency and induces an electromotive force, which can be measured by receiver coils, eventually producing time signals called free induction decays (FID). The magnetic behavior of the altered magnetization vector and consequently the generated contrast, can further be manipulated by using 90-, 180-degree, or other pulses and different application time periods.

If we consider the vectors in a standard Cartesian coordinate system, the rate that the atoms recover to 66% of the original M0 field along the z-axis is called the spin-lattice or T1 relaxation time. Moreover, the recovery rate along the xy-plane, which occurs concurrently to the T1 relaxation time is called the spin-spin relaxation time or T2 relaxation time. In general, the simplified MR signal that results in contrast is defined by equation 1.2:

$$S = PD\left(1 - e^{-TR/T1}\right)e^{-TE/T2}e^{-bD} \quad (1.2)$$

in equation 1.2, TR (time of repetition), TE (time of echo), and the "b" parameter can be controlled by pulse manipulation and other factors, whereas PD, T1, T2, and D are dependent on the imaged tissue and cannot be altered. The term "D" is related to the water diffusion profile of the structure. More information on this specific parameter is given in section 1.2.4. Consequently, the contrast mechanism of MR imaging is dictated by the aforementioned four terms (PD, T1, T2, and D) which themselves are influenced by factors such as TE and TR times.

In order to understand TE and TR times, the fundamental MRI concepts need to be discussed in more detail. As mentioned before, considering an arbitrary tissue segment, upon the application of a strong external magnetic field (B0), the majority of the proton atoms will align with the direction of the B0 field, whereas a certain number of them will align anti-parallel. The antiparallel atoms are considered high energy atoms while the parallel atoms are considered to be in a low energy state. Moreover, if the effect of all such protons are combined, the net magnetization vector will be aligned with the B0 field. This net magnetization vector is also called the longitudinal magnetizations.

If an RF pulse, strong enough to push 50% of the original parallel protons into the antiparallel state is applied, the longitudinal magnetization will be at zero because the opposing magnetic fields will cancel each other out along the Z-axis. Consequently, due to the properties of vectors, another magnetization vector called the transverse magnetization vector will now appear in the XY plane. However, if the RF pulse is removed, the transverse magnetization vector will lose its magnitude overtime, ultimately becoming zero while the longitudinal magnetization gain magnitude as the setup reverts to its baseline state. The T2-relaxation or spin-spin relaxation is related the change in the magnitude of the transverse magnetization vector. Conversely, the T1-relaxation or the spin-lattice relaxation is related to the magnitude gain in the longitudinal magnetization vector. In this context, TR is the time between the application of RF pulses, whereas TE is the time between the application of the RF pulse and the detected FID signal by the MRI machine. Both the T1 and T2- relaxation times are dependent on the type of tissue that is being imaged as well as TE/TR and thus due to these factors, we can accentuate differences in T1 and T2-relaxations to image different tissue type.

2.2.2 T1-weighted MRI

In the context of Parkinson's disease, MRI has gained considerable attention due to its ability to depict morphological abnormalities in the substantia nigra and basal ganglia. T1-weighted MRI sequences can display the macro-structural degeneration profile of different Parkinsonian syndromes. Given the atrophic nature of Parkinson's disease, this imaging protocol seems to be a viable choice for analyzing structural alterations of the brain. The usage of morphological parameters extracted from high-resolution T1-weighted MRI seem especially promising due to the modality's comparably high discriminative power and widespread availability. In recent years, numerous studies utilizing information derived from structural T1-weighted images have been conducted to differentiate PD sub-syndromes, which will be discussed in this section.

As previously mentioned in section 1.2.1, different brain tissue types such as white/gray matter or cereberospinal fluid and others have unique PD, T1, T2, and diffusion profiles. Therefore, depending on the region of interest, we might prefer certain modalities over others as they might provide a better differentiation. T1-weighted images, due to their high spatial resolution, are optimal for analyzing grey matter structural morphology and are characterized by short TE and TR times. In general, the longitudinal or spin-lattice relaxation time is related to how fast or slow hydrogen atoms within a region of interest undergo magnetization realignment after application of an RF pulse. Clearly, as structures contain different types of tissue, each with their own specific T1-relaxation time, the T1-weighted MRI uses these relaxation time differences to create contrast. For instance, fatty structures have a slower T1 time than fluids, therefore, they appear brighter on T1-weighted MRI images and ultimately result in contrast between the aforementioned structures.

T1-weighted MRI Studies in Parkinson's Disease

In terms of T1-weighted studies in PD, most studies have reported increased atrophy in PSP compared to PD. Messina et al.[24], for example, showed significant volume reductions in the cerebellum, putamen, thalamus, pallidum, brainstem, and hippocampus. Cerebellar gray matter and mesencephalon white matter loss have been reported by Focke et al.[25] in PSP compared to PD. Morphological differences, such as white/gray matter volume loss, cortical thickness, and regional brain surface area changes have been reported in PD vs. PSP.[26–28] In a previous morphological study by Gama et al.[29], volumetric changes in the midbrain and pons area were reported. Moreover, Price et al.[28] and Quattrone et al.[30] found midbrain and cerebral peduncles differences in PD compared to PSP. A voxel-based morphometry (VBM) study, which is a method often employed to analyze structural differences, by Menke et al.[31] revealed significant shape differences in the right pallidum in PD compared to healthy controls (HC). Another VBM study conducted by Focke et al.[25] reported significant cerebellar grey matter loss in PSP compared to PD. In general, due to the clinical acceptance of T1-weighted imaging in PD, many morphometric studies have investigated the structural differences between PD and PSP.

2.2.3 T2-weighted MRI

In T2-weighted imaging, the contrast is calculated mostly based on the T2 relaxation time. This imaging protocol can quantify a surrogate for brain iron deposition within different regions such as the brain[32] and liver.[33] This is an important factor for PD disease severity considering that abnormal iron accumulation has been confirmed in the substantia nigra in postmortem studies.[32] The T2-weighted MRI sequence is characterized by long TE and TR times and a bright contrast

for fat and fluids. However, transverse magnetization decays faster than what is expected from natural mechanisms previously outlined in section 1.2.1. This rate is defined as the effective T2 and is denoted by T2* (T2 star). The difference between T2 and T2* (T2*<T2) is mostly caused by inhomogeneities in the B0 magnetic field. In fact, the T2*-weighted imaging protocol is well suited to recognize localized magnetic inhomogeneities created by factors such as iron accumulation in hemorrhages or other conditions, where magnetic disturbance is present. The governing relationship between T2 and T2* is denoted by equation 1.3:

$$\frac{1}{T2^*} = \frac{1}{T2} + \frac{1}{T2'} \quad (1.3)$$

where T2' (T2 prime) represents the relaxation time attributed to field inhomogeneities caused by various factors, most prominently iron. The reciprocals of T2 and T2* (R2 = 1/T2 and R2* = 1/T2*) are called relaxation rates and are often used to provide an "indirect" marker for the quantification of brain iron accumulation levels. R2* (R2 star), R2, and R2' (R2 prime) are significantly correlated with iron concentration, therefore intensities of R2*, R2, and R2' images are often used as a surrogate marker of iron aggregation.[34] Figure 1.3 depicts structural T1-weighted images as well as quantitative R2 maps employed in this study from a healthy subject.

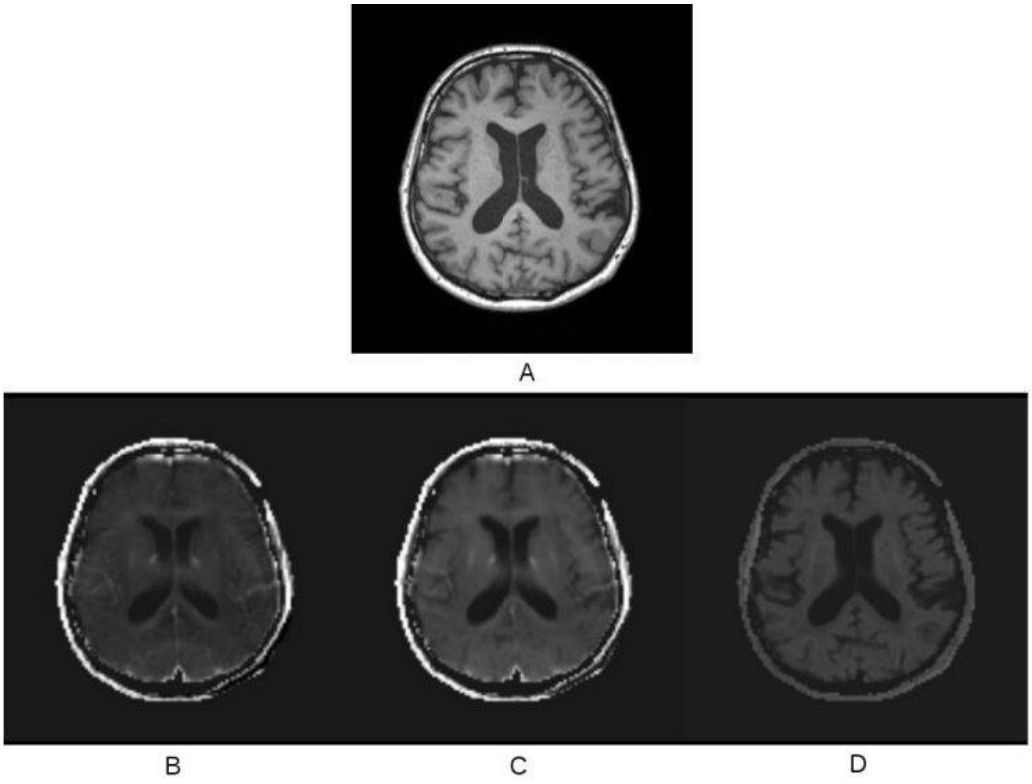

FIGURE 1.3. TRANSVERSAL VIEW OF T1/T2-WEIGHTED MRI IMAGES OF A HEALTHY SUBJECT (A) STRUCTURAL T1-WEIGHTED MRI, (B) QUANTITATIVE R2 (= 1/T2) MAP, (C) R2 PRIME, AND (D) R2 STAR MAPS

T2-weighted MRI Studies in Parkinson's Disease

In general, T2-weighted studies in PD are less prevalent than T1-weighted based studies. Wang et al.[35] reported significantly higher levels of iron accumulation in multiple system atrophy (MSA), which is another form of atypical Parkinsonian syndrome different than PSP-RS, in the putamen and pulvinar thalamus compared PD. Unlike PSP-RS, MSA is not at the focus of this research thesis and is only mentioned by name hereafter. Interestingly, a study by Lee et al.[36] showed a non-significant increase in putaminal R2* levels in PSP compared to PD. Similarly, a recent study by Du et al.[37] found significantly higher R2* values in substantia nigra and the red nucleus in PSP compared to PD. Following a longitudinal study conducted by Ulla et al.,[32] increased R2* levels were identified in the substantia nigra and caudal putamen in PD compared

to healthy controls (HC). Overall, multiple contradictory studies have been published regarding R2/R2* levels in PSP vs. PD that merit further investigation.

2.2.4 Diffusion-weighted MRI

Unlike T1- and T2-weighted MRI, diffusion-weighted imaging (DWI) is a MR imaging modality that measures the diffusion profile of brain regions. Unlike conventional T1-weighted MRI, micro-structural axonal changes following neurodegenerative disease can be observed using this imaging modality. It is worth mentioning that as of now, DWI imaging has no clinical utility in PD diagnosis. In this section, DWI fundamentals will be briefly described.

As mentioned before, MRI produces contrast via the manipulation of the proton nuclei in water molecules following the simplified equation of 1.2. However, in DWI, the contrast is related to the term D, which specifies the translational or Brownian motion of water particles. In order to produce a diffusion-weighted image, the coefficient "b" needs to be altered. Consequently, TE and TR are held constant while two experiments are performed with different "b" values. Following this process, the diffusion term can be calculated:[38]

$$D = (-ln\frac{S_2}{S_1})/(b_2 - b_1) \quad (1.4)$$

The term "b" is related to the application of different gradients via gradient coils, which was previously explained in section 1.2.1. In essence, two gradients are applied along a single axis, effectively altering the phase and consequently causing a phase difference of the water molecules within a voxel. If another gradient pulse along the same axis is applied with the opposite polarity, a phase refocusing is expected. However, refocusing is in fact marginal due to the translational

movement of the water molecules and other factors. In general, the "b" term can be altered by adjusting the strength and the timing of the gradient pulses. Finally, by calculating equation 1.4 in all voxels, the apparent diffusion coefficient (ADC), which is essentially the same as medium diffusivity (MD), in each voxel is obtained.

In DWI, water motion can only be measured along the applied gradient axis. However, if the effects of all gradients (X,Y,Z) within any given voxel are combined, ADC can be calculated along any orientation. This is an important feature of DWI because there is often clinical value in knowing the precise orientation of fibers within a specific region of interest. Fiber orientation, however, is not necessarily aligned with the main three axes and are in fact oblique to them. Consequently, to pinpoint the dominant tract orientation, diffusion tensor imaging (DTI) is employed. In DTI, measurements along the X, Y, and Z axes are mapped to a 3D ellipsoid, which denotes the average diffusion distance in each direction. The geometric information of this ellipsoid (also called eigenvalues) such as longest axis (Lambda1), middle axis (Lambda2), and shortest (Lambda3) axis can be used to calculate diffusion tensor properties such as fractional anisotropy (FA), medium diffusivity (MD, same as ADC), axial diffusivity (AD), and radial diffusivity(RD). Hypothetically, if the diffusion were to be isotropic then the ellipsoid would turn into a sphere meaning that:

$$\lambda 1 = \lambda 2 = \lambda 3 \quad (1.5)$$

Fractional Anisotropy (FA)

Fractional anisotropy (FA) measures the micro structural integrity of brain tissues and is a prominent biomarker used in neuroimaging studies.[39] Higher FA values indicate higher

anisotropy meaning that water molecules face aligned structures such as white matter fiber, membrane, and myelin. Fractional anisotropy is calculated by equation 1.6:

$$\mathrm{FA} = \sqrt{\frac{1}{2}} \times \left(\frac{\sqrt{(\lambda 1 - \lambda 2)^2 + (\lambda 2 - \lambda 3)^2 + (\lambda 3 - \lambda 1)^2}}{\sqrt{(\lambda 1 + \lambda 2 + \lambda 3)^2}} \right) (1.6)$$

Medium Diffusivity (MD)

In mean diffusivity (MD), which is another notation of ADC, values are inversely related to membrane density and are, thus, higher in the cerebrospinal fluid and damaged tissue structures.[40] MD is a sensitive marker of cellularity and is calculated by equation 1.7:

$$\mathrm{MD} = \frac{(\lambda 1 + \lambda 2 + \lambda 3)}{3} (1.7)$$

Axial Diffusivity (AD)

Axial diffusivity (AD) is related to white matter changes such as the aging process where white matter tracts have higher AD values.[41] AD is equal to the longest ellipsoid axis.

$$\mathrm{AD} = \lambda 1 \ (1.8)$$

Radial Diffusivity (RD)

Radial diffusivity (RD) is related to the changes in axonal diameter and density[41] and is calculated by equation 1.9:

$$\mathrm{RD} = \frac{(\lambda 2 + \lambda 3)}{2} (1.9)$$

Figure 1.4 depicts the various DWI maps used in this study.

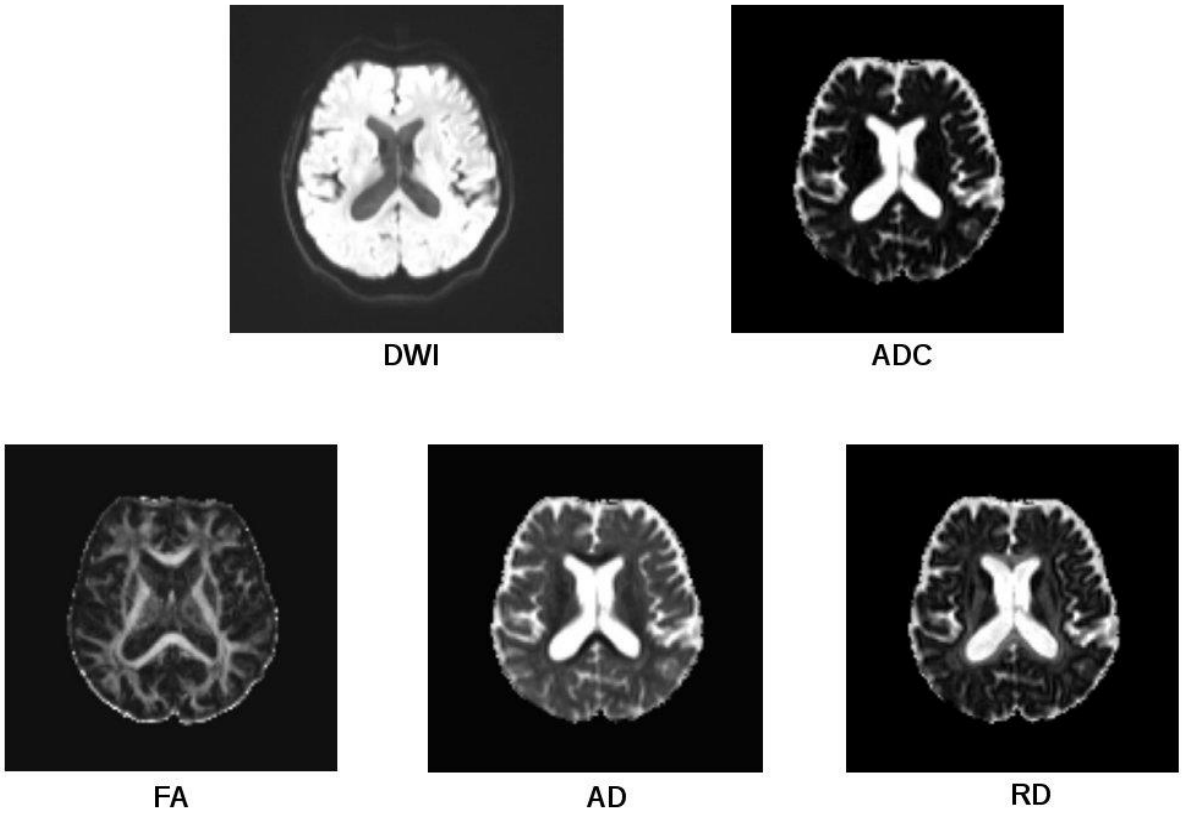

FIGURE 1.4. DIFFERENT DIFFUSION MAPS DERIVED FROM THE NATIVE DTI IMAGE. IMAGES SHOWN ARE FROM A HEALTHY SUBJECT INCLUDED IN THIS STUDY. AS MENTIONED BEFORE, ADC AND MD ARE OFTEN USED INTERCHANGEABLY.

DWI Studies in Parkinson's Disease

Nicoletti et al.[42] reported a significant increase of regional ADC values in the putamen, caudate, globus pallidus, thalamus, and the precentral white matter in PSP compared to PD. In two similar studies, higher ADC values were found in PSP subjects in the superior cerebellar peduncle compared to PD and healthy controls (HC).[43,44] The diffusion profile of the superior cerebellar

peduncles and corpus callosum have been found to be distinguishing factors in PSP and PD.[24,25] In another study, higher ADC values in globus pallidus and midbrain in PSP compared to PD were reported.[47] Furthermore, in line with the previous studies, increased ADC values in the putamen, globus pallidus, and caudate nucleus in PSP compared to PD were identified.[27] Moreover, differences in putamenial diffusivity and fractional anisotropy of substantia nigra were shown.[49] Gattellaro et al.[4] found that ADC values are increased in the substantia nigra, genu of the corpus callosum, and in the superior fasciculus in PD compared to HC. Furthermore, reduced FA values were found in the supplementary motor area, pre-supplementary motor area, and cingulum in PD compared to HC.[50] Lower FA values in PSP compared to HC in the frontol-orbital area, supplementary motor area, and other areas have been reported.[51] A recent study found that FA values in the olfactory area are highly discriminative for the differentiation of PD from HC.[52] Furthermore, increased AD and RD values in the substantia nigra, midbrain, and thalamus in PD compared to HC were previously shown.[53]

2.3 Data Processing

2.3.1 Image Registration

In order to extract the required parameters from the MRI modalities mentioned in the previous sections, the multi-channel datasets need to be processed and prepared in order to automatically extract quantitative imaging measurements, which can be used for a subsequent classification. Registration is defined as the determination of a geometrical transformation that aligns points in one image of an object (often called the moving image) with corresponding points in another image (often called the reference image) of the same object or another object. There are two

general types of transformation, linear and non-linear, with different degrees of freedom (DOF). The DOF is defined as the number of independent ways that a transformation, be it linear or non-linear, can be altered. Moreover, in medical image registration, a combination of transformations, also called concatenating, are employed to apply inter/intra modality normalization.

Rigid Transformation

The rigid transformation has a DOF of 6 consisting of 3 translations and 3 rotation terms. Therefore, the corresponding arrangement of data points are kept constant meaning that the shape and volume of an object is not altered following this type of registration

Affine Transformation

The affine transformation on the other hand, has 12 degrees of freedom with 3 terms for each of the translation, rotation, scaling, and skewness. This transformation, which is considered a more general form of the primitive rigid transformation, preserves distance ratios and maps lines to lines while not preserving actual angles and distances. Many of the intra-patient registration tasks in medical imaging are satisfied by a combination of rigid and affine transformations.

Non-linear Transformation

Taking a step further, any transformation with more than 12 degrees of freedom is considered a non-linear registration. The non-linear transformation approach covers image registration tasks that are not satisfied with the conventional linear rigid+affine registration model. In other words,

whenever the relationship between the points in images is non-linear, linear functions are not able to map the relationships accurately. Following the advent of medical imaging studies in recent years, numerous linear and non-linear registration tools have been developed with varying degrees of success.[54] In this research thesis, the advanced normalization toolkit[55](ANTs) was employed for the necessary registration tasks. The ANTs software package includes a state-of-the-art non-linear transformation called symmetric normalization (SyN).

Figure 1.5 depicts the results of a rigid, affine, and non-linear registration of the MNI152 atlas[56] to a T1-weighted image of a PSP-RS subject in a concatenated fashion. The MNI152 atlas is derived from 152 structural images that were averaged following a non-linear registration to the native MNI coordinate system. The registration quality is often checked purely by visual inspections. The primary goal here is to check whether the borders as well as anatomical locations such as the ventricles, brainstem, and others are registered reasonably well. Figure 1.6 shows a failed registration attempt where the ventricles are clearly mis-aligned. In case of figure 1.6, one must change the method or adjust the registration parameters for optimal results, which is often time consuming yet a crucial step in the entire registration process.

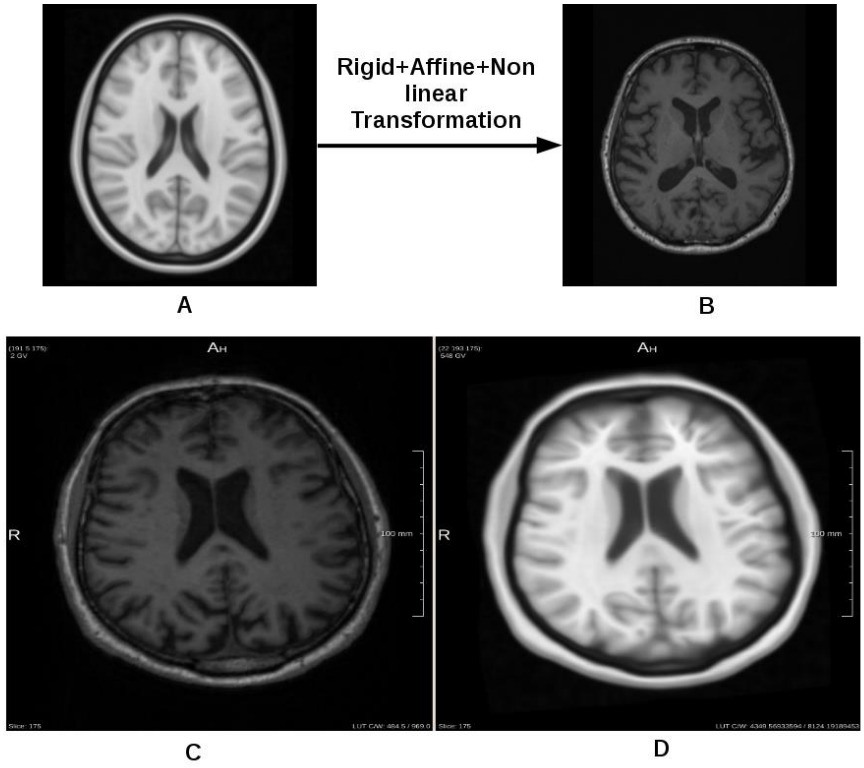

FIGURE 1.5. NON-LINEAR REGISTRATION OF (A) MNI ATLAS TO (B) T1-WEIGHTED MR IMAGE. RESULTS OF REGISTRATION DEPICTED IN (D) SHOWS SATISFACTORY ALIGNMENTS BETWEEN STRUCTURAL LANDMARKS SUCH AS VENTRICLES AND BRAIN BORDERS IN IMAGE (C). NOTE: (B) AND (C) ARE THE SAME PATIENT BUT AT A DIFFERENT SLICE NUMBER.

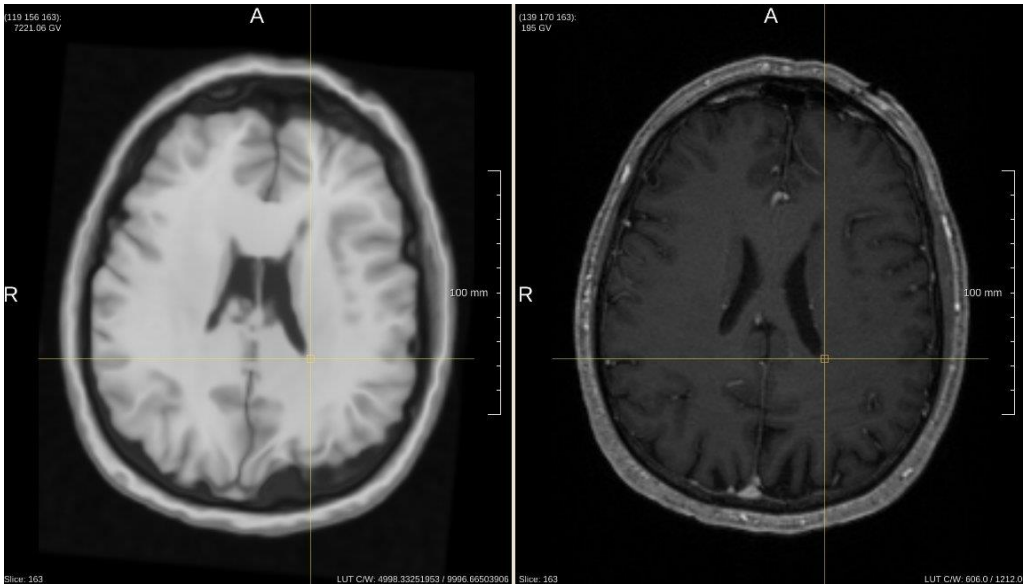

FIGURE 1.6. NON-LINEAR REGISTRATION OF MNI ATLAS TO T1-WEIGHTED MR IMAGE USING ANTS. RESULTS OF REGISTRATION (DEPICTED IN LEFT) SHOW MISALIGNED (SMASHED) VENTRICLE STRUCTURES AND ULTIMATELY SIGNIFIES A FAILED REGISTRATION ATTEMPT.

2.4 Feature Selection & Classification

2.4.1 Overview

Classification techniques or supervised machine learning methods, in the context of Parkinson's disease differentiation, utilize a set of continuous, categorical, or binary input features, to identify and differentiate PD syndromes through complex mathematical models. The aforementioned "learning" methods are conducted via a process called training. In machine learning, training is defined as exposing the model to labeled training datasets so that multi-dimensional pattern can be identified (learned), which can then be used to classify new unseen datasets based on the same features. For validation of the classification results, the full training set is typically split into a training and testing set such that the classification performance can be

determined using cross-validation principles. In this research project, regional brain information obtained from T1-, T2- and diffusion-weighted MRI are used as input features. As a primary step, many classification routines employ a feature selection method, which removes non-informative and redundant features, thus decreasing data dimensionality. After feature selection, the classifier is typically trained in the second step based on training data that contains a ground truth diagnosis for each case. Therefore, clinicians have to provide this ground truth classification, which can also be a problematic step since histo-pathologically proven diagnoses are typically not available in image-based PD studies. To overcome this drawback, the training set is typically limited to patients for whom a confident diagnosis can be secured.

2.4.2 Feature Selection Methods

Feature selection methods are commonly employed to reduce data dimensionality and eliminate redundant and non-informative features. In this work, a wide variety of feature selection methods such as correlation-based, information gain, principal component analysis, RELIEFF, and others are implemented.[57] A brief explanation of the feature selection methods used in this research project are described below. For more in-depth details regarding each feature selection method, the interested reader may refer to the provided reference.

Correlation-based Feature Selection

The goal of correlation-based feature selection is that the optimal feature subset will have features highly correlated with the class or label, yet uncorrelated with one another. Therefore, following a standardized Pearson correlation coefficient approach, the correlation feature selection algorithm ranks features based on their total numbers, intra-correlation, as well as their

correlation with the corresponding class. In this method, correlation between features and the output variable will be calculated and attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1) are selected.[58]

Gain Ratio & Information Gain-based Feature Selection

In the gain ratio method, which is based on decision tree learning methods, the classification significance of features is measured via the gain ratio with respect to the class. Similar to the gain ratio model, in the information gain-based method, the entropy for each attribute for the output variable can be calculated. Entropy values vary from 0 (no information) to 1 (maximum information). Those attributes that contribute more information will have a higher information gain value and can be selected, whereas those that do not add much information will have a lower score and can be removed.[59]

Principal Component Analysis (PCA)

In Principal Component Analysis (PCA), a number of eigenvectors are selected to account for 95% of the variance in the initial dataset.[60] In fact, the selected eigenvectors will replace the original feature set by effectively creating new features that are based upon linear combinations (e.g. weighted averages) of several features of the original feature space. In detail, the diagonal correlation matrix is calculated and the intra correlations of features are determined. Here, a positive correlation close to +1 indicates high correlation.[61]

RELIEFF

In the RELIEFF feature selection method, a random feature [F] is selected, then the [K] nearest features from the same class (called hits) as well as the [M] nearest features from the opposite class (called misses) are selected. This method then iteratively updates the weight of the features based on a mathematical equation described by Kononenko et al.[62], for a total of [N] number of times, where ultimately highly weighted features denote more relevance. Due to the weighting equation, RELIEFF is effective in classification problems where features exhibit strong linear and non-linear dependencies.[63]

SVM-based Feature Selection

In binary support vector machines (SVM), the decision boundary, which is often called hyperplane, is optimized in a way to maximize the margins between the two classes. The margins are defined as the distance between the closest class instance and the boundary hyperplane. Moreover, in case of a linear kernel SVM, the slope of the linear decision boundary function is called the weight vector, which includes a combination of the individual weights corresponding to all the training instances. In fact, only the weights belonging to training cases located at borderline regions of the boundary are non-zero, therefore contributing the most to the overall weight vector of the decision boundary function. The SVM-based feature selection method utilizes the magnitude of the aforementioned weight vector as an indicator of feature relevance. In the first step, an SVM is trained based on the complete feature set of the training examples and their corresponding class labels. Consequently, the magnitude of all weight vectors is calculated and the feature with the least amount of significance is omitted from the feature list.

In the next round, the complete feature list minus the aforementioned feature are used to repeat the process. The feature ranking is concluded when all the features are listed from most to least relevant.[64]

2.4.3 Classification Methods

The next stage after the preprocessing steps is to train the classifier with the feature set. As three classes will be classified in this research project, only inherently multilevel classification methods were employed here. Conceptually speaking, multi-level classifiers can be separated into the following categories (1) inherently multi-class (2) multi-class as one vs. one, or (3) multi-class as one vs. all. In the first category, the classification is performed in a pure multi-level paradigm, whereas in the second and third category the classification is divided into separate binary problems and the results are averaged across the investigated binary classifications. Commonly used inherent multi-level (as opposed to inherently binary classifiers) classification methods include support vector machines (SVM), k-nearest neighbors (K-NN), random forest (RF), and others. A short description of the aforementioned classifiers is provided below. Similar to the feature selection steps described in section 1.4.2, the reader is encouraged to refer to the provided references for in-depth description of the mathematical concepts behind the classification algorithms.

Decision Trees

Figure 1.7 represents an example of using decision trees to classify a given binary situation (i.e. playing golf or not playing golf). Here, the decision tree is constructed based on training examples that contain categorical feature information about the weather. Consequently, the

features are used to split the training dataset into decision subsets or decision nodes. In detail, we know that the outlook feature has three possible variations, namely sunny, overcast, and rainy. In the next step, we look at each of the leaves and decide which ones are "pure" or "un-pure". Based on the labels of the training datasets in figure 1.7, "pure" leaves are the ones when the individual either plays or does not play. In this example, whenever the outlook is overcast, we see that the individual does indeed play golf. If a pure leaf is reached, the algorithm ignores them and moves on to other un-pure leaves. In this example, the "sunny" and "rainy" leaves are considered un-pure and need to be further split to be considered pure. In the next stage, we split the "windy" feature following a sunny outlook as it provides the best target label separability. Based on the training dataset, it seems that whenever it is windy the individual does not play and vice versa. We further apply the same principle to the humidity attribute and construct the decision tree. By building a model like this, one is able to classify new datasets. The way the features are split is based on the theory of entropy (not always the case) where the model prefers a maximum number of pure leaves following the initial data splitting round.[65] Decision trees have a high interpretability compared to other classification methods, detect the most discriminatory features, and other significant advantages.[66]

FIGURE 1.7. CONSTRUCTION OF A DECISION TREE USING WEATHER-BASED FEATURES (LEFT) IN ORDER TO PREDICT IF A PERSON WILL PLAY GOLF OR NOT. (HTTP://WWW.SAEDSAYAD.COM/DECISION_TREE.HTM)

Random Forrest

Random forest (RF) is a supervised ensemble decision tree classifier, meaning that it consists of several uncorrelated decision trees. In detail, as described before, decision trees classify instances by sorting based on feature values and creating decision leaves. In a random forest approach, multiple "random" decision trees (a total number of K trees) are created. The training examples are randomly separated into K different sets where each individual tree will learn that specific training set that it was assigned. However, this process is not conducted on the entire set of features, but a random feature set is composed for each tree. The final prediction of RF is typically decided based on a majority voting scheme following the results of each of the K decision trees. Random forest classifiers have several important benefits such as being intuitive and easy to use and implement and consistent high levels of accuracy.[67]

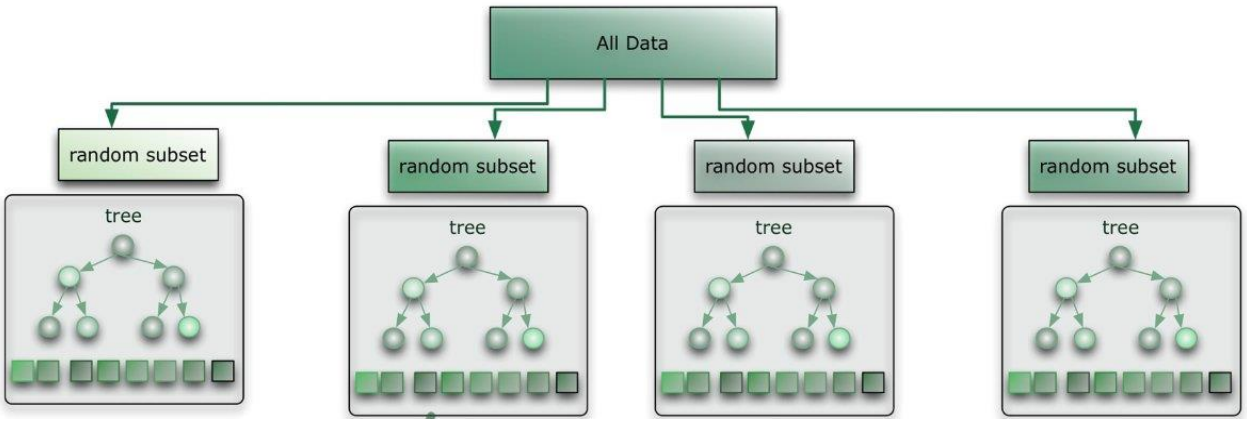

FIGURE 1.8. CONSTRUCTION OF A RANDOM FOREST USING VARIOUS DECISION TREES IN ORDER TO CLASSIFY INSTANCES BASED ON A MAJORITY VOTE SCHEME. (HTTP://BLOG.CITIZENNET.COM/BLOG/2012/11/10/RANDOM-FORESTS-ENSEMBLES-AND-PERFORMANCE-METRICS)

Logistic Model Tree

The logistic model tree (LMT) classifier combines logistic regression with decision trees. In order to understand the concept of logistic regression, one must first understand what linear/multiple regression entails. In a linear regression model, we map a number of data points with 'N" number of "continuous" features in space and intend on fitting a straight line that ultimately predicts the target value based on all the other features. Figure 1.9 depicts the linear relationship between high school and university GPA, where the data points are obtained from several students. This approach has numerous benefits mainly, that we are able to observe correlations between features (i.e. how does high school GPA correlate with university GPA). However, in a logistic regression model, as the name suggests, we are interested in predicting a binary instance (0 or 1) rather than continuous variables (such as GPA scores). Furthermore, instead of fitting a line, we fit a sigmoid-shaped logistic function between 0 to 1. Consequently, we may calculate probabilities of class labels based on where new data points fall with respect to

the curve. In LMT, the nodes (leaves) of the tree consist of logistic regression models instead of fixed values as described in the decision tree section. The models in each node can be refined continuously at higher level trees, effectively outperforming decision trees.[68]

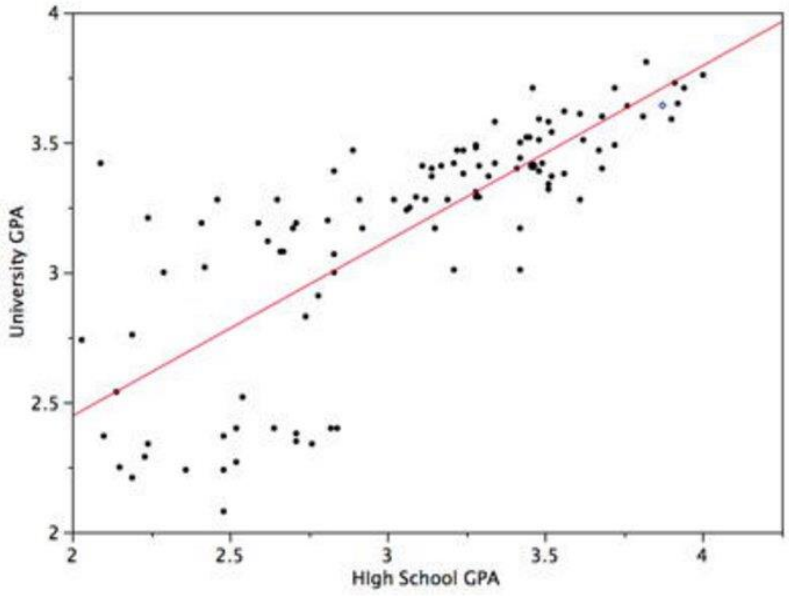

FIGURE 1.9 LINEAR RELATIONSHIP BETWEEN HIGH SCHOOL AND UNIVERSITY GPA DERIVED USING A LINEAR REGRESSION MODEL. (SOURCE: HTTP://ONLINESTATBOOK.COM/2/REGRESSION/INTRO.HTML)

K-Nearest Neighbor

The k-nearest neighbor (K-NN) algorithm is a "lazy" learning approach, meaning that it classifies new instances locally based on the most similar cases in the training set. In balanced datasets, an object is typically classified by a majority vote of its neighbors, with the object being assigned to the class that is most common amongst its "k" nearest neighbors (e.g. based on Euclidean distance or other metrics). Consequently, in a binary classification task, in order to avoid a voting tie situation, the "k" must be (1) odd and (2) not be a multiple of the number of

classes. In a broader sense, instances can be considered as points within an n-dimensional instance space where each of the n dimensions corresponds to one of the n features that are used to describe an instance. In figure 1.10, depending on K, the unknown data point (star) could belong to class 2 (if k=3) or class 1 (if k=5). Several benefits of this method are high flexibility, minimum operational cost during training and overall simplicity.[69]

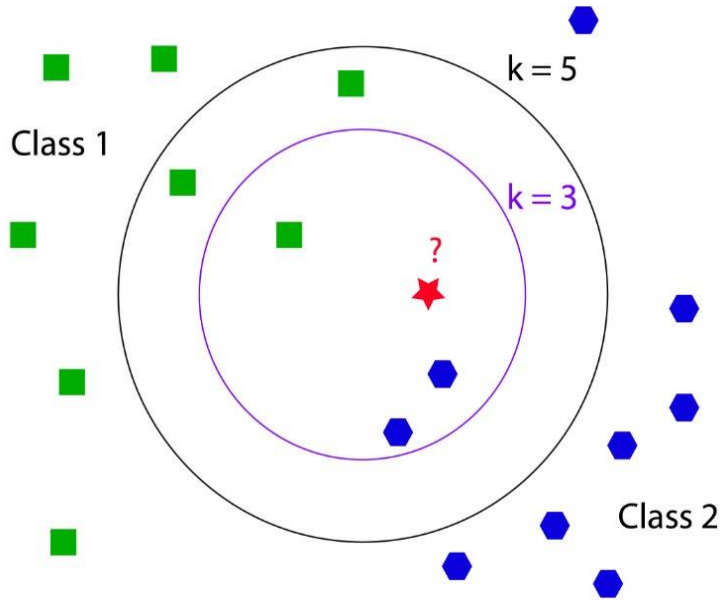

FIGURE 1.10. SCHEMATIC REPRESENTATION OF A KNN CLASSIFIER. (SOURCE: HTTP://WWW.COXDOCS.ORG/LIB/EXE/FETCH.PHP?MEDIA=PERSEUS:USER:ACTIVITIES:MATRIXPROCESSING:LEARNING:KNN.PNG)

Naive Bayes

The Naive Bayes (NB) classification method is based on the combination of likelihood prediction and prior probability-based concepts. In figure 1.11, a set of training data points (either green or red) are scattered across space where the number of green instances is twice that

of red data points. According to Bayesian analysis, an unknown test data point (here shown in white and denoted as X) is twice as likely to have a green membership than red. Therefore, considering a total of 60 training data points, the prior probability for green and red membership is 40/60 and 20/60, respectively. In order to classify X, we draw a circle around it with an arbitrary radius (chosen as a classification parameter before training) and calculate the number of different data point within the circle. Consequently, the likelihood of X given green or X given red is 1/40 and 3/20, respectively. Clearly, as the number of red data points in the vicinity of X is larger than green, X is more likely to belong to the red class, even though prior probability assumed a green membership. The NB classification method ultimately combines the effects of prior probability and likelihood to assign class labels to unseen data points. Based on Bayes rule, the posterior probability of X being green or red is 1/60 and 1/20, respectively, which results in a red membership assignment for X.

NB has three major types such as multinomial, Bernoulli, and Gaussian. The NB sub-type for a given problem is chosen depending on the type of features used in the corresponding classification task. Discrete frequency features such as words in a text file, binary features, and normally distributed features require the multinomial, Bernoulli, and Gaussian, respectively. However, all subtypes follow the same initial assumption, where the feature values are independent of one another and that all features contribute to the probability of the predicted class label in some fashion. In a Gaussian NB, we assume the features for each class have a Gaussian distribution. During the training phase, we calculate the mean and variance of each feature per case and build a probability model. New unseen datasets are given class/label probabilities based on the relation of their features and the established model. The main

advantage of NB is that the training datasets do not necessarily have to be large for a comprehensive model.[70]

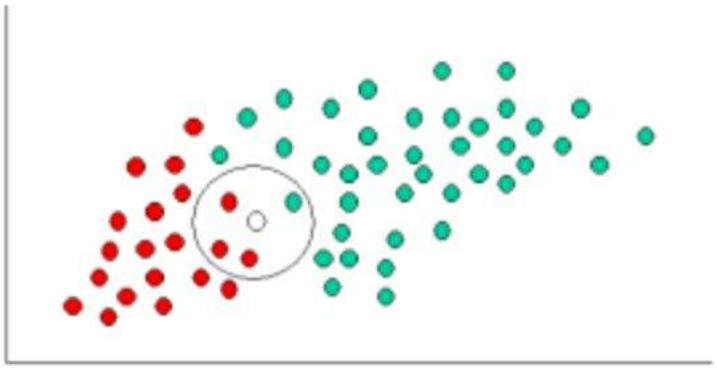

FIGURE 1.11. SCHEMATIC REPRESENTATION OF A NB CLASSIFIER. (SOURCE: HTTP://WWW.STATSOFT.COM/TEXTBOOK/NAIVE-BAYES-CLASSIFIER)

Support Vector Machines

Support vector machines (SVM) use an n-dimensional hyperplane to optimally separate the dataset into two or more data classes.[31] In a binary classification, a margin is assigned on either side of the hyperplane that separates the datasets. Maximizing this margin creates the largest possible distance between the hyperplane and the instances, which in turn reduces the generalization error via a so-called quadratic optimization problem with a set number of constraints. In a multiclass classification, however, constraints are added for each class and thus the size of the quadratic optimization is proportionally expanded. Figure 1.12 represents a schematic representation of a linear SVMs and the obtained margin. The model complexity of SVM is not affected by the number of cases in the training set, meaning that SVM is best used for datasets with large number of features. However, many classification problems are non-linear

therefore, no linear separation can be constructed. A common solution for this problem is to map the data into a higher dimensional space and define the hyperplane accordingly via kernel functions. This new transformed space is called the transformed feature space. Several kernel types have been proposed for SVM such as linear, radial-basis-function (RBF), and polynomial kernels. The kernel function should be carefully chosen since it is directly related to the transformed feature space in which the training instances will be classified.[71] Consequently, unlike the linear kernel, as RBF and polynomial kernels can be highly optimized to fit the dataset, proper consideration should be taken to avoid an over-fitted model (refer to section 1.4.4 for an explanation on over-fitting and why it should be avoided). SVMs have several important benefits, namely consistent high performance as well as efficiency as the algorithm only uses a small subset of training points for establishing the hyperplane.

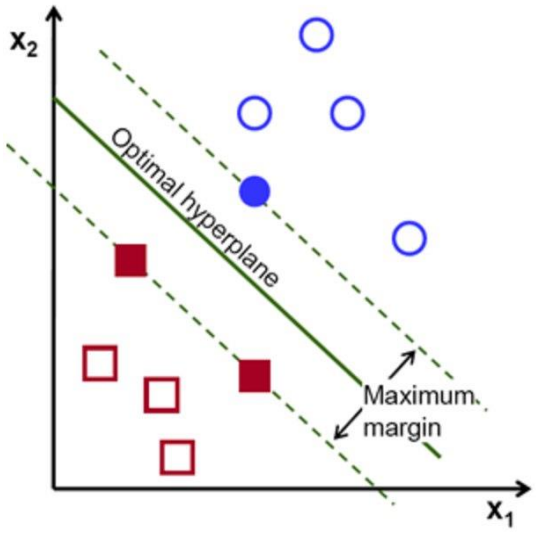

FIGURE 1.12. SCHEMATIC REPRESENTATION OF AN SVM CLASSIFIER WITH A MAXIMUM MARGIN.
(SOURCE:HTTPS://DOCS.OPENCV.ORG/3.0-BETA/DOC/TUTORIALS/ML/INTRODUCTION_TO_SVM/INTRODUCTION_TO_SVM.HTML)

Multi-layer Perceptrons

Multi-layer perceptrons (MLP) or neural networks consist of several "layers", namely, input, hidden, and output layers. Figure 1.13 depicts a representation of a standard MLP architecture. Each layer consists of multiple nodes (perceptrons), where the nodes themselves are primitive functions such as linear combination functions, step functions, and others. The response (or output) of each node is the combination of features (also called weights and shown as "w" in figure 1.13) that it receives in the input layer. The input layer consists of all the features corresponding to the classification problem. Moreover, the input layer is the input to the first hidden layer, and the input of the second hidden layer is the output of the first hidden layer and so on. Ultimately, this complex web of perceptrons converge on the output layer where the classification takes place. The connection between nodes is defined by their weights that are iteratively optimized through a back-propagation approach in the training stage. MLP can be used both in regression and classification tasks and they have a significant ability to (1) model non-linear and complex functions (2) generalize and many other capabilities.[72]

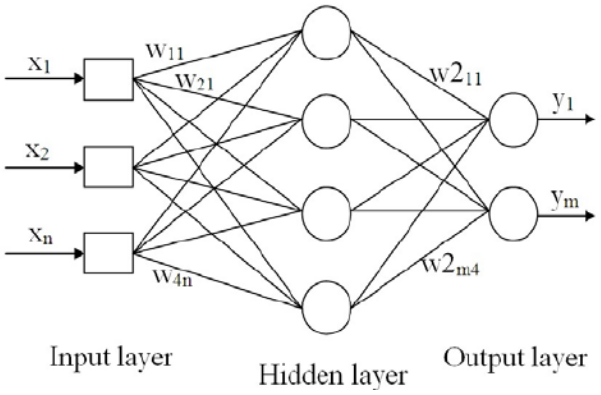

FIGURE 1.13. SCHEMATIC REPRESENTATION OF A MLP CLASSIFIER CONSISSTING OF INPUT, HIDDEN, AND OUTPUT LAYERS. (SOURCE: HTTPS://WWW.RESEARCHGATE.NET/FIGURE/A-SCHEMATIC-DIAGRAM-OF-A-MULTI-LAYER-PERCEPTRON-MLP-NEURAL-NETWORK_FIG3_257071174)

2.4.4 Classification Validation Methods

There are several standard metrics, which can be used to evaluate the performance of classifiers following a k-fold cross validation. In a k-fold cross validation scheme, the full sample size is randomly separated into k equal sized sub populations. Essentially, out of the k sub populations, one of them is labeled as validation data and the rest of the k-1 is used as training data for classification. Following cross-validation procedures, several important evaluation metrics are calculated namely, sensitivity, specificity, accuracy, kappa statistics/error (KS), precision, F-measure, Matthews correlation coefficient (MCC), and the area under the receiver operating characteristics (AUC). These metrics will be discussed in this section.

Many classification problems in medical applications are primarily focused on binary or three level (two or three case) differentiation. For example, in this project, a subject either belongs to the PD, PSP-RS, or healthy controls (HC) group, where the overall objective of the classifier is to place each subject in the correct group. Here, the "correct" group is defined based on the real clinical diagnosis of each subject, which is often called the gold standard. The outcomes of such classification routines are defined in statistical terms such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN) and are represented in the so-called confusion matrix (figure 1.14 represents a binary confusion matrix).

Disease Condition	Predicted as Yes	Predicted as No
Actual: Yes	TP	FP
Actual: No	FN	TN

Figure 1.14. Conceptual representation of a binary confusion matrix

Considering a binary classification between PD and HC, sensitivity or true positive rate (TPR) is defined as TP divided by the summation of TP and FN whereas, specificity or true negative rate (TNR) is defined as TN divided by the summation of TN and FP as shown below.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (1.10)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (1.11)$$

In detail, sensitivity depicts the percentage of subjects with a specific PD syndrome who have been correctly identified as such by the classifier. Specificity denotes the percentage of HC who have been correctly grouped as not PD by the classifier, meaning that HCs are correctly identified as not having Parkinson's disease. Classification accuracy is defined by the summation of TP and TN divided by all the possible outcomes as shown below:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (1.12)$$

In many "balanced" classification tasks where the number of subjects in each group are reasonably similar, this metric serves as a reliable indicator of classification performance.

However, in many real-life classification problems, the number of subjects in each group is not completely balanced therefore, additional evaluation metrics need to be employed to compare classifiers. One of these measures is the Matthews correlation coefficient (MCC) or the phi-coefficient, which is used in classification tasks to denote a correlation coefficient between the obtained classification and the ground truth.[72] For example, if two classifiers result in the same accuracy but have different MCC values, the classifier with the highest MCC (<1) is often regarded as the better option between the two. MCC is described by equation 1.13:

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad (1.13)$$

Receiver operating characteristic (ROC) curves are plotted using TPR versus false positive rate, as the discrimination threshold of the classification method is changed. Consequently, the area under the curve (AUC), which has a value from 0.5 to 1, evaluates the performance of a classification, where 1 denotes an excellent classifier and 0.5 shows classification at chance level. Furthermore, Kappa statistics/error (KS), which is used to measure inter-rater agreement (i.e. classification accuracy vs. chance level accuracy) and compare the performances of classifiers, is a well-suited method to analyze classifications based on chance (look at equation 1.14)

$$\mathrm{KS} = 1 - \frac{1-\mathrm{acuuracy}}{1-\mathrm{Pc}} \quad (1.14)$$

Here, the term "Pc" is defined as the probability of chance agreement, which would be at 50% considering a completely balanced dataset.

The F-measure (also called F1 score) is defined as the harmonic mean of sensitivity and precision, whereas precision is defined as TP divided by the summation of TP and FP. Equations 1.15 and 1.16 represent precision and F1 score. In this research project, the aforementioned metrics will be used to evaluate classifier performance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1.15)$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision.Sensitivity}}{\text{Precision+Sensitivity}} \quad (1.16)$$

Another important evaluation metric is the permutation test, which can be used to validate that the obtained results are in fact statistically significant and not simply a result of noise. In short, after an optimal classification model is determined, it will be tested against various permutations of the initial dataset. This is achieved by randomly reassigning the class labels to different data instances and testing the results with the optimal model. It is worth mentioning that the vast majority of existing CAD-PD research failed to implement a permutation test in their study.

Over-fitting/under-fitting in classification methods

As described above, the performance of a classification algorithm is heavily reliant on the training phase. Consequently, extra precautions need to be taken to avoid over-fitting or under-fitting of the model. Over-fitting is referred to instances where the classification model perfectly fits the training datasets, therefore compromising the overall generalizability with regards to other unseen datasets. Over-fitting is commonly encountered in non-linear models, where the classification algorithm is flexible enough to over learn the details of the training datasets. Common ways to avoid over-fitting are adding more datasets, cross validation, feature reduction,

the addition of noise and others. Conversely, depending on how the training procedure is executed, under-fitting might occur, where the model fails to model the training datasets and therefore fails in the testing phase as well. Common methods to mitigate this situation is the optimization of classification parameters and investigation of other classification algorithms.

2.5 State of the Art

In the following sections, a comprehensive modality-specific overview of studies, which have used MRI-based parameters and features for PD classification tasks are presented. In each of the papers reviewed, several key information such as the type of MRI protocol used, subject population, extracted features, and most importantly, in-depth information on the classifier itself namely, classifier type, whether a feature selection step was involved, and evaluation metrics like sensitivity, specificity, and accuracy among other information are disclosed.

2.5.1 Classification Studies based on T1-weighted Images

Based on section 1.2.2, it was described that morphological features which essentially measure structural atrophy in Parkinsonian syndromes are important classification features. In this section, studies that have used these features such as volume, brain surface area, and others are presented.

Scherfler et al.[73] proposed a classification model based on a decision tree classifier to differentiate PSP, MSA, and PD subjects based on volumetric features from T1-weighted MRI scans. All images collected were analyzed using FreeSurfer[74] software to gain volume measurements of 22 sub-cortical brain regions. A portion of the dataset was split into a training

set applied to the decision tree algorithm and was able to find the most discriminative volumetric feature regions. The overall data set utilized in this study consisted of 40 PD, 40 MSA, and 30 PSP subjects, and the classification method was evaluated by comparing results with a final clinical diagnosis gold standard. By only using training datasets, the diagnostic accuracy of this method was found to be 97.4% for the case of differentiating between PD and MSA or PSP.

Sarica et al.[75] described a rather simple classification method to differentiate HC, PD, and PSP. In this study, the volume of 15 brain regions, such as subcortical structures, ventricles, white matter, and the cortex, were automatically determined after segmentation using the Freesurfer software and used for classification. Numerous classification methods such as naive Bayes, random forest, and SVM, were compared using a 10-fold cross validation scheme. The evaluation cohort of 46 HC, 65 PD, 32 PSP subjects revealed that the SVM classifier led to the best accuracy of 58.56% for differentiating HC from PD subjects, while the naive Bayes performed best for differentiating HC vs. PSP (93.75% accuracy) and PSP vs. PD (95% accuracy).

Instead of extracting the volume from T1-weighted datasets for selected brain regions, Duchesne et al.[26] proposed a more advanced method to automatically differentiate HC, PD, and atypical Parkinson syndromes including PSP and MSA subjects. After preprocessing, the T1-weighted dataset is non-linearly registered to a brain atlas. After registration, the intensity and Jacobian of the non-linear deformation field are extracted for each voxel within a volume-of-interest in the atlas that encompasses the complete hindbrain region. Positive Jacobian determinants indicate a volume increase while negative values indicate volume decrease. Principal component analysis

(PCA) was used to reduce the dimensionality of the feature space. After this, an SVM with least square optimization was used for group separation in the multi-dimensional PCA space. Leave-one-out evaluation based on 149 HC, 16 PD, and 16 subjects diagnosed with either PSP (n=8) or MSA (n=8) revealed that the proposed method achieved an overall classification accuracy of 90.66%, a mean sensitivity of 93.22%, and mean specificity of 88.2%. No information is given regarding the class-specific accuracies achieved by this model.

Salvatore et al.[76] presented a morphology-based classification method to differentiate PD, PSP, and HC that is similar to the method described by Duchesne et al. Briefly described, the T1-weighted MRI datasets were non-linearly registered to the MNI brain atlas[56] after preprocessing. In contrast to Duchesne et al., the Jacobian determinant was not used for the classification and the volume-of interest was not restricted to the hindbrain region. However, a PCA was also used to extract the voxel-based features from the T1-weighted MRI datasets and a linear SVM was used for the final classification task. The proposed method was evaluated using datasets of 28 PD, 28 PSP and 28 HC. The leave-one-out cross validation resulted in a mean accuracy for PD vs. control, PSP vs. control and PSP vs. PD of 85.8%, 89.1%, and 88.9%, respectively.

Rana et al.[77] presented another method to classify PD and HC using T1-weighted MRI datasets based on an SVM classifier. The T1-weighted datasets are spatially normalized and tissue probability maps of gray matter, white matter, and cerebrospinal fluid are automatically generated using the statistical parametric mapping (SPM) software[78] and smoothed using a Gaussian filter. After this step, the substantia nigra, thalamus, hippocampus, frontal lobe, and midbrain are automatically segmented for each subject using an atlas-based approach and the

three tissue probabilities are extracted for each voxel and each segmented brain structure. A feature selection based on the mutual information metric is applied to reduce the number of features for each brain region, which are then used for the classification. The proposed method was evaluated based on 30 HC and 30 PD subjects using a leave-one-out cross validation scheme. Overall, the best classification results were achieved using only information about gray matter probabilities in the substantia nigra resulting in an accuracy of 86.67%, sensitivity of 90%, and specificity of 83.3%. However, the exact motivation for the proposed method remains unclear since structures such as thalamus, hippocampus, and the substantia nigra belong to the cerebral gray matter and should not contain white matter or cerebrospinal fluid without the smoothing applied.

More recently, Singh and Samavedham[79] proposed a method to automatically differentiate HC and PD subjects. After skull-stripping, spatial normalization, and registration to the MNI brain atlas, the gray and white matter was automatically segmented using the SPM software. After this, the image intensity values are extracted for each subject separately for the white matter and gray matter segmentations. In addition, a Kohonen self-organizing map[80] was used for feature selection for the white and gray matter features, whereas the resulting set of parameters was then used as the input for a SVM classification with a radial basis function kernel. Using the study cohort of 245 HC and 518 PD, the 10-fold cross validation showed that this method achieves an accuracy of 95.04% for the complete patient sample. For classifying PD vs. HC, certain GM regions such as medial dorsal nucleus, putamen, pulvinar, Brodmann areas 29 and 23 and for WM, corpus callosum and limbic cortex were found to be especially relevant. PD vs. HC comparison revealed that lateral posterior and anterior nuclei of thalamus were affected most.

Focke et al.[25] utilized T1-weighted images to classify PD, MSA, PSP and HC by investigating the volume loss in different brain regions in a voxel-based morphometry study. Image segmentation was conducted using the SPM software. In this voxel-based SVM classification, local weightings were imposed on the maps to include anatomical information about different patient groups. They separated the classification into two subgroups of white matter only and grey matter only analysis and calculated the classifier's accuracy with and without local weights. In this context, using a study cohort of 21 PD, 11 MSA, 10 PSP and 22 HC, the highest accuracy using leave one out cross validation was 96.8% for PSP vs. PD when local weights were imposed. In MSA vs. PD the classifier's accuracy was 71.9% when local weights were involved. Moreover, the classifier was not able to distinguish between HC and PD since the obtained accuracy were all less than 50%, which is interestingly lower than the accuracy expected by chance. Based on this VBM analysis, PSP and MSA were associated with white matter loss in the mesencephalon and putaminal grey matter loss, respectively. They also found that there is a cerebellar grey matter loss in PSP compared to PD.

Peng et al.[81] conducted a study using a multi-kernel SVM classifier to differentiate between HC and PD subjects, based on two types of features generated from T1-weighted MRI datasets. Features were extracted using BrainLab software and included up to 390 low level ROI features such as GM volume and cortical surface area, as well as 3003 high level correlative features. Feature selection for classification was a 3-step process starting with a t-test followed by a minimum redundancy maximum relevance step, ending with an SVM-based recursive feature elimination resulting in 15 discriminative features. This method was evaluated using a dataset of 69 PD and 103 HC subjects, and utilized a two-layer nested cross validation scheme both

consisting of 10-fold cross validation to validate the classifier. The classification accuracy was calculated at 85.78% whereas, sensitivity and specificity were 87.64% and 87.79 respectively.

Mueller et al.[82] presented a study in which they compared the performance of SVM classifiers (implementing either a linear or polynomial kernel) in order to classify PSP subjects from HC through VBM using T1-weighted MRI datasets. Images were processed utilizing SPM and Matlab. Two approaches were employed for feature extraction, where the first method was voxel-based generation of gray matter probability maps and the second was based on disease specific ROIs. All methods proposed in this study were evaluated using a study cohort of 20 PSP and 20 HC subjects and a leave-one-out cross validation scheme. The highest results were from the use of a polynomial kernel combined with ROI feature selection presenting an accuracy of 85%. It is worth noting, that the datasets used for this study were acquired from four different scanner types with differing imaging parameters.

Lin et al.[83] conducted a study where different binary multiple logistic regression classifiers were compared in their ability to identify PD from HC through the use of volumetric features. Images were preprocessed using the VBM8 software and segmented into WM, GM and CSF, then aligned to the MNI space using DARTEL algebra registration.[54] Composite anatomical atlases were then created using Hammer's maximum probability atlases.[84] For feature selection, a two-sample student's t-test was implemented to test for differentiation significance. Features with a p-values less than 0.1 were selected. To assess the classifiers performance, a dataset of 72 PD and 73 HC subjects was used as well as a leave-one-out cross validation scheme for each

classifier. The highest performing classifier was able to achieve an accuracy, sensitivity, and specificity of 74%, 74%, and 75%, respectively.

2.5.2 Classification Studies based on T2-weighted Images

PD classification employing T1-weighted features were extensively investigated by previous studies, whereas, T2-weighted imaging studies in PD have been used comparably less for classification. In this section, an overview of T2-weighted studies and similar brain iron accumulation measuring protocols such as magnetization transfer imaging (MTI) and susceptibility-weighted MR imaging (SWI) are described.

Boelmans et al.[85] proposed a linear discriminant method to classify PD, PSP, and HC subjects using quantitative T2' values determined from triple-echo T2 and T2* MRI sequences. Therefore, average T2' values were measured in multiple deep gray and white matter structures. Training and testing of the stepwise linear discriminant analysis classifier using the average T2' values in the caudate, pallidum, putamen, and thalamus as input features was conducted using a database of 30 PD, 12 PSP, and 24 HC subjects. The results revealed that this method is capable of classifying subjects to the three groups with an overall accuracy of 74.2%. Of the 24 HC subjects, 18 were classified correctly whereas 6 subjects were grouped in the PD group. In the clinically diagnosed PD group, 9 and 2 subjects were wrongly grouped as HC and PSP, respectively. All PSP subjects were classified correctly. Moreover, statistical analysis of the T2' values in the analyzed brain regions identified shortened T2' values in caudate nucleus, globus pallidus and putamen in PSP compared to HC and PD.

Eckert et al.[86] used a stepwise linear discriminant analysis using indirect brain iron content measures obtained from globus pallidus, putamen, substantia nigra, caudate nucleus, and white matter to classify HC, PD, MSA, and PSP subjects. Image segmentation of the aforementioned structures was carried out manually. Using a study cohort of 20 HC, 15 PD, 12 MSA and 10 PSP, out of the 20 HC, 5 subjects were wrongly classified as PD. In 15 PD subjects, 3 subjects were incorrectly grouped as HC. In case of 10 PSP subjects, only one subject was classified as MSA. However, for the 12 MSA subjects, only 7 subjects were correctly classified as MSA, whereas 4 and 1 subjects were falsely classified as PSP and HC respectively.

Haller et al.[87] presented another voxel-based SVM classification method based on T2-weighted images to differentiate PD and "other" subjects. The "other" group consisted of 4 MSA, one PSP, and other atypical variant of PD. Briefly described, the images were linearly registered to the MNI brain atlas and the iron measurement surrogate signal values werenormalized using the average signal value of the ventricular system. Each brain tissue voxel with the normalized values was considered as a feature for the classification using an SVM with a radial basis function kernel. The most discriminative 100, 250, 500, 750 and 1000 features as identified by the RELIEFF algorithm were used independently in training and testing of the SVM classification models. As for evaluation purposes, a 10-fold cross validation using 16 PD and 26 "other" subjects revealed that the SVM model trained with the 100 most discriminative yields the highest average accuracy of 86.92%. Furthermore, using these number of features resulted in a sensitivity and specificity of 87%.

2.5.3 Classification Studies based on DTI Datasets

Information extracted from diffusion-tensor MRI (DTI) has been found especially advantageous for examining white matter integrity in various neurological diseases and may identify potential differences at a microstructural level in Parkinsonian syndromes.[88] Consequently, as microstructural changes are typically expected to precede macrostructural (volumetric) changes, DTI might indicate brain abnormalities at an earlier stage than structural T1-weighted images. The most relevant quantitative DTI parameters are apparent diffusion coefficient (ADC), which measures the degree of tissue water diffusivity, fractional anisotropy (FA), an indicator for axonal integrity, radial diffusivity (RD), which is associated with white matter myelin, and axial diffusivity (AD), which provides a metric for axonal degeneration.[41] The typical fingerprint of degenerated neuronal tissue is an increase of ADC, RD, and AD but a decrease of FA.[4,44]

Scherfler et al.[89] proposed a simple threshold-based method to classify PD and HC subjects based on ADC values in the olfactory tracts. In brief, ADC parameter maps were calculated from DWI datasets and registered to the MNI brain atlas. To discriminate between PD and HC, voxel clusters of the olfactory tracts that showed the strongest ADC differences between the two groups were identified using voxel-wise t-tests to determine a threshold to separate the two groups ($0.78 * 10^{-3}$ mm^2/s). Twelve HC and 12 PD subjects not part of the test set were used for this cluster and threshold analysis. For classification of new subjects, these clusters are registered into patient space and the corresponding patient-specific mean ADC value is extracted and employed for subsequent classification using the ADC threshold. The proposed method was evaluated using a test set of 9 PD and 8 HC subjects resulting in an accuracy of 94.1% with one HC subject being incorrectly classified as PD.

Salamanca et al.[90] proposed a method to classify PD and HC subjects using a combination of a fisher vectors[91] and logistic linear regression based on FA and ADC values in 14 regions of interest as features. The MNI brain atlas was non-linearly registered to the DTI data and the resulting non-linear transformation field was used to warp 14 regions of interest from the MNI atlas to the patient space to extract the corresponding FA and ADC values. Using a database of 100 subjects (50 PD and 50 HC), the proposed method was evaluated using 100 experiments of 10-fold cross-validation resulting in a maximal accuracy of 77%. No information regarding the specificity or sensitivity is described.

Haller et al.[92] presented an approach to classify PD and subjects with atypical forms of Parkinsonism using an SVM classifier and voxel-wise FA values as features. Briefly, FA parameter maps were generated from DTI data and non-linearly registered to an average FA brain atlas. After this step, each voxel of the brain tissue was used as a feature for the classification using an SVM with a radial basis kernel. To reduce the feature space, a feature selection was performed using the RELIEFF[36] algorithm extracting the most discriminative 100, 250, 500, 750 and 1000 features (voxels), which were then separately used for the SVM classification. Each SVM classification model with the varying number of training features was evaluated using a 10-fold cross validation. Using a study cohort of 17 PD and 23 atypical forms of Parkinsonism, followed by a 10-fold cross validation routine, the best classification result with an accuracy of 97.5% was achieved for the SVM using the top 100 features. Voxel-based statistics revealed a significant increase in FA and a decrease in ADC and RD in the right frontal WM in PD compared to HC. Furthermore, using the top 100 features resulted in a sensitivity and specificity of 94%, and 100%, respectively.

Banerjee et al.[93] utilized an SVM classifier along with a radial basis kernel to differentiate between PD and HC subjects via features based on Cauchy deformation tensors (CDT)[94] extracted from diffusion MRI. The ensemble average propagator biomarker includes information about the shape and direction of diffusion processes at the individual voxel level and is used to create CDT features for morphometric analysis. FA values were also calculated in the same ROI to compare classifier performance in CDT only feature set versus an FA only feature set. A PCA based algorithm known as principal geodesic analysis (PGA)[95] was utilized to reduce the dimensionality of the feature set. The proposed method was evaluated using a study cohort of 46 PD and 22 HC subjects using a leave-one-out cross validation scheme. In case of using FA features only, the classification accuracy was 76.47%, whereas sensitivity and specificity were calculated at 78% and 73%, respectively. However, the proposed CDT features were found to perform better than FA features leading to a binary classification accuracy of 98.53%, sensitivity of 98%, and specificity of 100%. The authors point out that the superior performance of CDT features may be because they contain diffusion direction and shape information, whereas FA maps show shape information only.

2.5.4 Classification Studies based on Multi-modal MRI Datasets

In the previous sections, research papers focused on individual level classification of Parkinsonian syndromes and healthy controls using features derived from single modality MRI were described. In fact, most of existing research in this field relies on single channel parameters. However, a few research papers have combined information from multiple sources such as diffusion and T1-weighted MRI in the hopes of building more comprehensive classification models, which are described below.

Péran et al.[96] presented a multi-parametric logistic regression method to classify PD and HC using image-based features of iron deposition, atrophy, and micro-structural damage. In this study, the thalamus, putamen, caudate, pallidum, substantia nigra, and red nucleus were semi-automatically segmented in high-resolution T1-weighted MRI dataset employing the FSL software. Moreover, FSL was used to analyze the T2*-weighted and DTI MRI datasets. More precisely, the DTI datasets were used to calculate parametric maps of fractional anisotropy and mean diffusivity, while the multi-echo T2*-weighted datasets were used to calculate quantitative R2* maps. All three parametric maps were then registered to the T1-weighted dataset to calculate average values for each parameter and brain region. Additionally, the volume of each structure was also calculated. Training and testing of the logistic regression was conducted using a database of 30 PD, and 22 HC subjects and a 10-fold cross validation. It was found that the combination of mean R2* values in the left or right substantia nigra, FA in the right substantia nigra and ADC in the putamen or caudate nucleus leads to the best discriminating power resulting in a global accuracy of more than 95%. No further classification was performed by using single modality features. An additional voxel-based analysis revealed that PD subjects exhibited higher R_2^* values in substantia nigra, lower FA in substantia nigra and thalamus and higher ADC in thalamus compared to HC, while no volumetric differences were found.

Morisi et al.[97] proposed a linear kernel SVM model with a feature ranking mechanism to classify PD, PSP, and MSA subjects based on features derived from DTI data, proton spectroscopy MRI, and morphometric volumetric analysis. Therefore, mean FA and ADC were determined in multiple brain regions, while other features were calculated using a histogram–analysis method or semi-automated volumetric analysis. More specific details regarding the image analysis

procedure are not described by the authors. After extraction of 152 features, a relative entropy method was used to rank the most differentiating features for each of the two-class classification problem. The proposed method was evaluated using a database consisting of 6 PD, 17 PSP, 15 MSA subjects and leave-one-out cross validation. In the case of one class vs all, the classification accuracies for PD vs all, PSP vs all, and MSA vs all were measured at 90%, 95%, 93%, respectively.

Du et al.[37] utilized an elastic-net regularized regression approach in order to differentiate between PD, MSA, PSP, and HC subjects using region of interest features based on DTI data and apparent transverse relaxation rate (R2*). FA and ADC maps were generated using DTIPrep[98], and R2* maps were determined using a Matlab tool developed for this study. For feature selection, the Bonferroni method was implemented to correct for multiple comparisons of MR images and the elastic-net model[95] was utilized with a nested 10-fold cross validation scheme to reduce the high dimensionality of the classification. The prospective method was tested using datasets of 16 MSA, 19 PSP, 35 PD, and 36 HC subjects following a 10-fold cross validation scheme. All binary classification accuracies obtained by the combination of DTI and R2* features were higher than 90%. Out of the 6 binary classifications performed, the best performance calculated was PD vs. PSP using both DTI and R2* data calculating a sensitivity of 97% and a specificity of 100%. It is worth noting that the combination of dual modality features resulted in higher accuracies compared to single modality features.

Planetta et al.[99] proposed a method applying an SVM classifier in combination with an ROI-based ROC analysis, which was then followed by a forward feature selection method to identify

differentiating features between PD, MSA, PSP, and HC subjects based on diffusion MRI data in combination with clinical test scores. Diffusion MRI data was preprocessed using FMRIB software and a bi-tensor model was generated, which calculated free water values at the voxel level and free water corrected fractional anisotropy values (FAt), which were ultimately registered to MNI space. Consequently, a total of 19 ROIs, including the anterior substantia nigra, globus pallidus, and others were considered as input features to the classification. Clinical test scores were also collected from multiple tests such as the Montreal Cognitive Assessment (MoCA)[100] and MDS-UPDRS-III. This classification method was assessed using a study cohort of 18 PD, 18 MSA, 18 PSP, and 18 HC subjects following a 10-fold cross validation scheme. In short, both diffusion and clinical measurements were used separately as well as in combination with each other with the SVM classifier. Overall, 20 binary combinations were reported in this study. Based on the results, most of the binary classification metrics showed improvements when diffusion and clinical features were combined. Out of all the classification experiments performed, PD vs. PSP resulted in the best performance reaching 100% sensitivity and 100% specificity.

Cherubini et al.[101] combined whole brain VBM and DTI measurements on 21 PSP and 57 probable PD subjects. Volumes were analyzed by FMRIB software whereas the nonlinear transformation calculated from the T1-weighted volume to the standard space was combined with the affine transformation calculated from the mean diffusivity (MD) and fractional anisotropy (FA) maps to the T1-weighted space, thereby aligning all individual data to the standard space. Each voxel was assigned 4 locally weighted features, such as gray matter atrophy, white matter atrophy, MD and FA, which were then used in the classifier. Following a

leave one out strategy using the SVM classifier (without any feature extraction) step was used for classification. The classifier reached 100% accuracy when only white matter atrophy was considered. However, when FA, MD and gray matter parameters were used without the white matter predictor, the sensitivity and specificity dropped to 90% and 96%, respectively.

Nemmi et al.[102] proposed a classification method to differentiate PD and HC subjects based on gray matter density, subcortical nuclei volume, and brain region shapes. Segmentation of the anatomical brain regions was conducted using FIRST[103] for volumetric analysis of subcortical nuclei such as amygdala, caudate nucleus, hippocampus, globus pallidus, putamen, left and right nucleus accumbens, and thalamus. Moreover, DTI was used to analyze structural connectivity distribution between local atrophy and cortical regions. Linear discriminant analysis was performed on the two groups using both global volume and shape only differences (i.e. gray matter density, subcortical nuclei volume, and subcortical nuclei shape). Linear discriminant analysis with leave-one-out cross validation method was used for (1) several brain region shapes (2) global volume combinations such as left putamen, left caudate or left caudate and right putamen, and (3) brain region shapes to evaluate classification accuracy on 21 PD and 20 HC. For the volume only analysis, the classifier reached accuracies up to 65% for both the left putamen and the combination of left caudate and left putamen. In case of shape only analysis, the reported accuracies were up to 83% for the left putamen and also the combination of left caudate and left putamen. The study concluded that subcortical shape results (local atrophy-only) were able to better categorize PD and controls compared to standard volumetric-only analysis.

A study conducted by Ota et al.[104] focused on classification of HC, MSA, and PD subjects. Information derived from DTI (such as fractional anisotropy), through tract based spatial

statistics, and gray matter volume data obtained from SPM were used as classification feature in three different types of discriminant functions. Accuracies of discriminant functions obtained from (1) stepwise methods (using 4th ventricle volume, substantia nigra, superior temporal region (ST), prefrontal region (PF)) and (2) two independent variables (pons, ST) and (3) Three independent variables (4th ventricle volume, middle cerebellar peduncle (MCP) and ST) were compared. Using a study cohort of 21 HC, 30 MSA, and 21 PD, a maximum accuracy of 88.9% was reported for both the stepwise and three independent variables methods.

The study by Long et al.[105] followed a multi-model approach by using functional and structural data to classify PD and HC. For the functional section, resting state fMRI (rsfMRI) images were analyzed in three categories such as amplitude of low frequency fluctuations (ALFF), regional homogeneity (ReHo), and regional functional connectivity strength (RFCS). In this context, ALFF, ReHo and RFCS maps were partitioned into 116 ROIs using the anatomical labeling atlas (AAL).[106] Moreover, structural information such as volume data from white matter, gray matter, and cerebrospinal fluid were analyzed whereas the segmentation was done using the default tissue probability maps. An SVM classifier with a hyperbolic tangent function as the kernel with a leave-one-out cross validation was used to determine the classification accuracy. The feature selection phase was conducted by utilizing a two-sample t-test to rank the most discriminative feature between HC and PD. Following a leave one out cross validation on 19 PD and 27 HC subjects, the classifier obtained a maximum accuracy of 86.96% when both structural and functional features were used. However, the classifier performed less favorable when functional and structural were used separately. For example, an accuracy of 73.91% was obtained when only ReHo, ALFF, and RFCS were used as metrics. Also, when ALFF and RFCS were used, an

accuracy of 67.39% was reached. However, when only structural metrics such as GM, WM, and CSF were involved the classifier obtained an accuracy of 80.43%. Furthermore, from the functional aspect, the authors found that PD subjects exhibited significant ReHo decrease in the bilateral ORBmid and ALFF decrease in the left ROL along with RFCS increase in the left PHG and ANG and right MTG compared to HC. Moreover, the PD group showed a volume decrease in the left PCL whereas a significant increase in the left PreCG and bilateral PCG was observed. The study finally concludes that multi-model approaches will lead to a better classification accuracy compared to single models.

CHAPTER THREE: OBJECTIVES AND HYPHOTESIS

As mentioned previously, individual level PD classification studies are currently limited to the combination of features from one, two or at the most three imaging modalities. In other words, the potential benefits of combining features from multiple sources has not been fully explored yet. According to existing literature, it is suggested that the combination of features obtained from various imaging modalities has the potential of resulting in higher classification accuracies.[35,96,99–101] Furthermore, most of the studies reviewed in section 1.5, had significant limitations, namely the usage of small study cohorts, non-rigorous validation routines such as 10-fold cross validation in the case of small sample sizes, lack of classification significance validation methods such as permutation tests, lack of a comprehensive feature selection approach, implementation of classification models prone to over-fitting, manual region of interest definition, and potential double dipping[107] scenarios in feature selection and classifications steps. Consequently, the obtained results might be overly optimistic and thus far from real clinical applications. A higher differentiation performance is highly critical for CAD-PD devices from multiple perspectives. Most importantly, while no medical interventions are currently available for PSP-RS, a reliable and robust classification of PD vs. PSP-RS will promote the integration of such devices within clinical frameworks and will, therefore, assist clinicians in the differential diagnosis of PD and PSP-RS. Generally speaking, advances in the field of computer science, machine learning in particular, have established the foundation for CAD methods to be adopted in real life clinical settings to assist in the diagnosis of a wide variety of diseases not limited to the brain. CAD device's numerous benefits are: (1) automation,

(2) consistent high levels of disease differentiation performance, (3) availability in remote areas where specialized clinicians are not always available, and (4) relative inexpensiveness.

Therefore, the aim of this research project is to present a framework of a CAD system for the automatic differentiation of HC, PD, and PSP-RS, called CAD-PD, which mitigates some of the limitations of previous methods described above. In order to establish this CAD-PD system, imaging information from various MRI modalities will be combined. In detail, multi-channel MRI datasets such as T1-weighted, T2-weighted, and diffusion tensor images are utilized to build a much more comprehensive CAD-PD compared to previously existing methods. As described in section 1.5, the theme of classifying PD sub-syndromes using imaging modalities has been explored in the past by different research entities, however none of them are as comprehensive as what will be presented in this study. Here, comprehensive is defined as follows:

1) Multi-modal feature sets are combined in the classification routine to elucidate potential benefits of each MRI modality in HC vs. PD vs. PSP-RS differentiation.

2) In this study, a wide variety of feature selection and classification methods are employed in a nested fashion to prevent double dipping. Unlike previous studies, multiple feature selection algorithms are implemented to identify the best performing method for PD sub-syndrome classification.

3) Rigorous classification validation methods are employed to reasonably evaluate the generalizability of the classification. Most previous studies thus far have had significant shortcomings in the validation of their classification pipelines.

4) A relatively large study cohort has been employed in the classification process so that the results are as general as possible. Many studies in the past have resorted to small study cohorts and are therefore less reliable for real life clinical implementation.

Ultimately, the primary hypothesis of this research endeavor is based upon the multi-model approach that is proposed. It is hypothesized that the combination of imaging features from multiple sources will improve HC vs. PD vs. PSP-RS classification performance compared to the usage of single modality features.

CHAPTER FOUR: MATERIALS AND METHODS

4.1 Study Cohort

The study cohort used for this work has been previously described in detail by Boelmans et al.[85]. Thirty-eight HC, 52 PD, and 21 PSP-RS subjects were scanned at the University Medical Center Hamburg-Eppendorf, Germany, using a 3T Siemens Skyra MR scanner. The clinical diagnosis of PD and PSP-RS was conducted according to the UK Brain Bank criteria[18,108] and the National Institute of Neurological Disease and Stroke[109], respectively. The inclusion criteria for the PSP-RS group were probable PSP-RS subjects presenting as classical Richardson syndrome with vertical palsy, axial rigidity, and balance instability with early falls. PSP-RS patients who exhibited prominent freezing phenomenon, asymmetric clinical features, and a clinically relevant levodopa response were excluded from the study. Prior to the study, informed consent was attained from all subjects. The study was approved by the local ethics committee. Table 3.1. represents an overview of the participating PD and PSP-RS subjects in the study.

Among others, the imaging protocol contained a high-resolution T1-weighted MPRAGE dataset and a DTI acquisition protocol. The high-resolution T1-weighted MPRAGE dataset was acquired using TR = 1900 ms, TE = 2.46 ms, flip angle = 90°, TI = 900 ms, image in-plane resolution of 0.94 mm^2, and 0.94 mm slice thickness. The T2-weighted and $T2^*$-weighted image sequences were also acquired for this study. In detail, for T2 determination, a spin echo with 15 echoes per shot was employed to record images at three different echo times of 12, 85, and 158ms in 74 seconds of acquisition. Moreover, the total number of slices was 24 with a thickness of 5mm as well as a field of view of 240mm, repetition time of 4590ms and a flip angle of 150 degrees. The

$T2^*$-weighted images were however, acquired via a single shot echo-planar sequence at a TE of 21, 52, and 88ms. The DTI sequence was acquired using a single-shot balanced echo-planar imaging sequence with TR = 4500ms, TE = 83ms, and flip angle = 90°. The DTI sequence consists of 27 contiguous transverse slices with a slice thickness of 5mm and in-plane resolution of $1.875mm^2$ acquired without diffusion gradients ($b=0s/mm^2$) and with diffusion gradients ($b=1000s/mm^2$) applied along 20 non-collinear directions, averaged over two acquisitions.

Table 3.1. Demographic and clinical characteristics of study participants

	Parkinson's disease	**Progressive supranuclear palsy – Richardson's syndrome**	**p-value**
No.	52	21	
Gender, F/M	17/37	11/10	p=0.093
Age at examination, y, mean ± SD (range)	65.5±8.6 (40-77)	71.1±5.5 (59-79)	p=0.104
Disease duration, y, mean ± SD (range)	12.7±6.8 (0.5-30.2)	5.9±3.3 (1.2-12.6)	p=0.001
Hoehn&Yahr, mean ± SD (range)	2.5±0.8 (1-4)	2.5±0.8 (1-4)	p=0.823
UPDRS motor score (OFF condition), mean ± SD (range)	36.8±13.1 (14-63)	32.7±11.7 (9-52)	p=0.658
UPDRS motor score (ON condition), mean ± SD (range)	19.9±10.2 (5-52)	28.7±10.6 (6-48)	p=0.003
MMSE, mean ± SD (range)	28.1±1.4 (23-30)	25.1±2.7 (19-29)	p=0.032

Due to missing image sequences (T1-, T2-, or diffusion-weighted) in 7 of the PD subjects and one of the PSP-RS subjects, those patients were excluded from the analysis. Therefore, 38 HC, 45 PD, and 20 PSP-RS subjects are available for this study.

4.2 Materials

4.2.1 Data Preparation

In this research work, different software tools and frameworks have been employed. An overview of these tools and a brief introduction of them is presented in this section. An important data preparation step is to convert the MRI images from the native proprietary scanner (Siemens Skyra MR scanner in this study) format, DICOM (dcm), to NIFTI fomat (nii). Simplicity, ease of use, conserved spatial orientation information are a few of the benefits of NIFTI over the DICOM image format. Moreover, NIFTI is easily handled in popular neuroimaging software tools available, making it a popular choice in the brain imaging community.

In this work, MevisLab and the dcm2nii conversion toolkit, which is distributed with the MRIcron package, were used for general format conversion. Mevislab can also be used for viewing, analyzing images, and even some basic image processing. The R2, R2*, and R2' parameter maps were obtained from the original T2- and T2*-weighted datasets using AnToNIa developed by Forkert et al.[110] Moreover, the DTI images in this study were further preprocessed using the DTI-Preprocess toolkit. This toolbox was used to perform minor distortion correction as well as generating the individual ADC, FA, RD, and AD maps.

4.2.2 Image Registration

Image registration was performed using NiftyReg and the advanced normalization toolkit[55] (ANTs) depending on the registration task performed. The difference between these tools is that ANTs has a more comprehensive and complex registration framework, allowing users multiple options and customization for each particular task compared to NiftyReg. However, while NiftyReg has fewer options than ANTs, it has been optimized to produce consistently desirable registration results for intra-modality registration tasks. Specifically, it was found that NiftyReg outperforms ANTS in the registration of atlases to individual T1-weighted images. Alternatively, ANTs was used when an inter-modality registration such as diffusion-weighted to T1-weighted MRI was considered. In detail, the automatic brain parcellation, which is required for the subsequent regional brain analysis, was performed by registration of the 152 MNI brain atlas[111] to each patient dataset. For this purpose, an affine registration of the atlas to the patient dataset was performed first using the block-matching approach described by Ourselin et al.[112] The resulting affine transformation was then used as an initialization for subsequent non-linear registration using a free-form deformation as implemented in the NiftyReg software package.[113]

Registration for extraction of T1-weighted MRI based features

The calculated non-linear transformation for each patient was used to warp the Harvard-Oxford subcortical, cortical, and the MNI regions to each patient employing a nearest-neighbor interpolation. These brain regions offer a fine parcellation of the brain into anatomical and functional regions while the single regions are still large enough to calculate proper quantitative parameters. More precisely, the Harvard-Oxford subcortical atlas defined in the MNI reference

space consists of 21 brain regions such as the thalamus, caudate, hippocampus, and brainstem and the Harvard-Oxford cortical atlas consists of 48 brain regions such as the insular cortex, precentral gyrus, and temporal pole. Finally, the MNI brain regions (not to be confused with the MNI 152 atlas used for registration) consist of 9 well known structures such as the caudate, cerebellum, frontal lobe, and others.

In addition to the parcellated atlas brain regions, the non-linear deformation field was also used for transforming a binary segmentation of the total intracranial volume to each patient dataset, which was used for volumetric normalization of the single brain regions to account for differences regarding the general head anatomy. Apart from the volume, the segmented brain regions were also used to determine the surface area of each brain region as well as the surface-area-to-volume ratio (SA:V). For calculation of the SA:V, the raw regional volumes instead of the volumes corrected for the full intracranial volume were used. Mevislab was used to check registration quality following the principles described in section 1.3.1 (see figure 1.6). Figure 3.1 depict the registration pipeline for the extraction of morphological features.

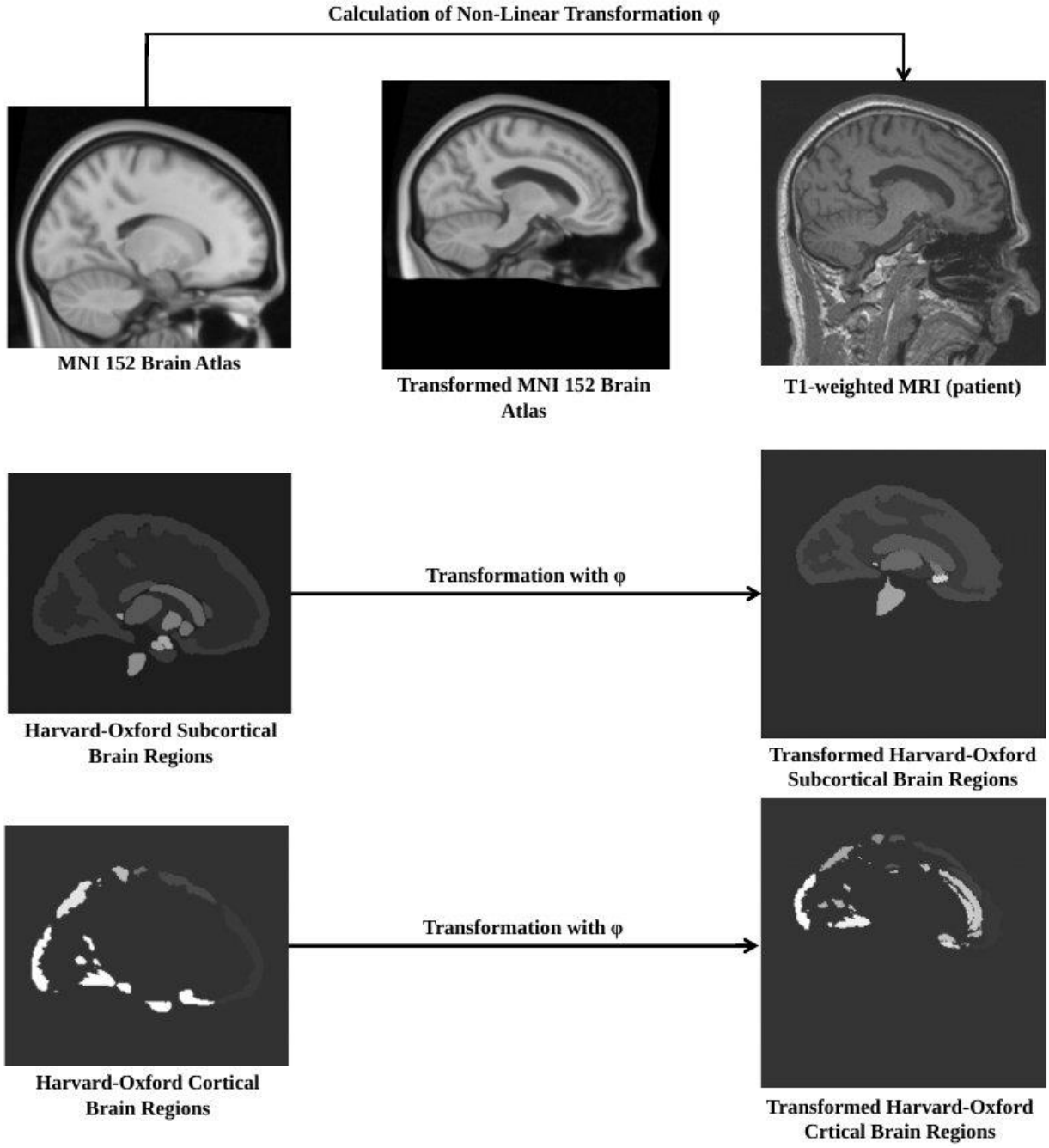

FIGURE 3.1. NON-LINEAR REGISTRATION OF MNI ATLAS TO T1-WEIGHTED MR IMAGE USING NIFTYREG FOR THE EXTRACTION OF MORPHOLOGICAL FEATURES SUCH AS VOLUME, BRAIN SURFACE AREA, AND SURFACE AREA TO VOLUME RATIO. THE REGISTRATION USING THE MNI REGIONS ARE NOT DEPICTED HERE.

Registration for extraction of T2 and diffusion-weighted MRI based features

Unlike the T1-weighted images, the T2-weighted and diffusion-weighted images were registered using ANTs. In detail, with respect to the extraction of diffusion-based features, the DTI sequence was registered to the T1-weighted MPRAGE image. The image processing pipeline for extraction of the regional ADC, FA, RD, and AD values is illustrated in Figure 3.2.

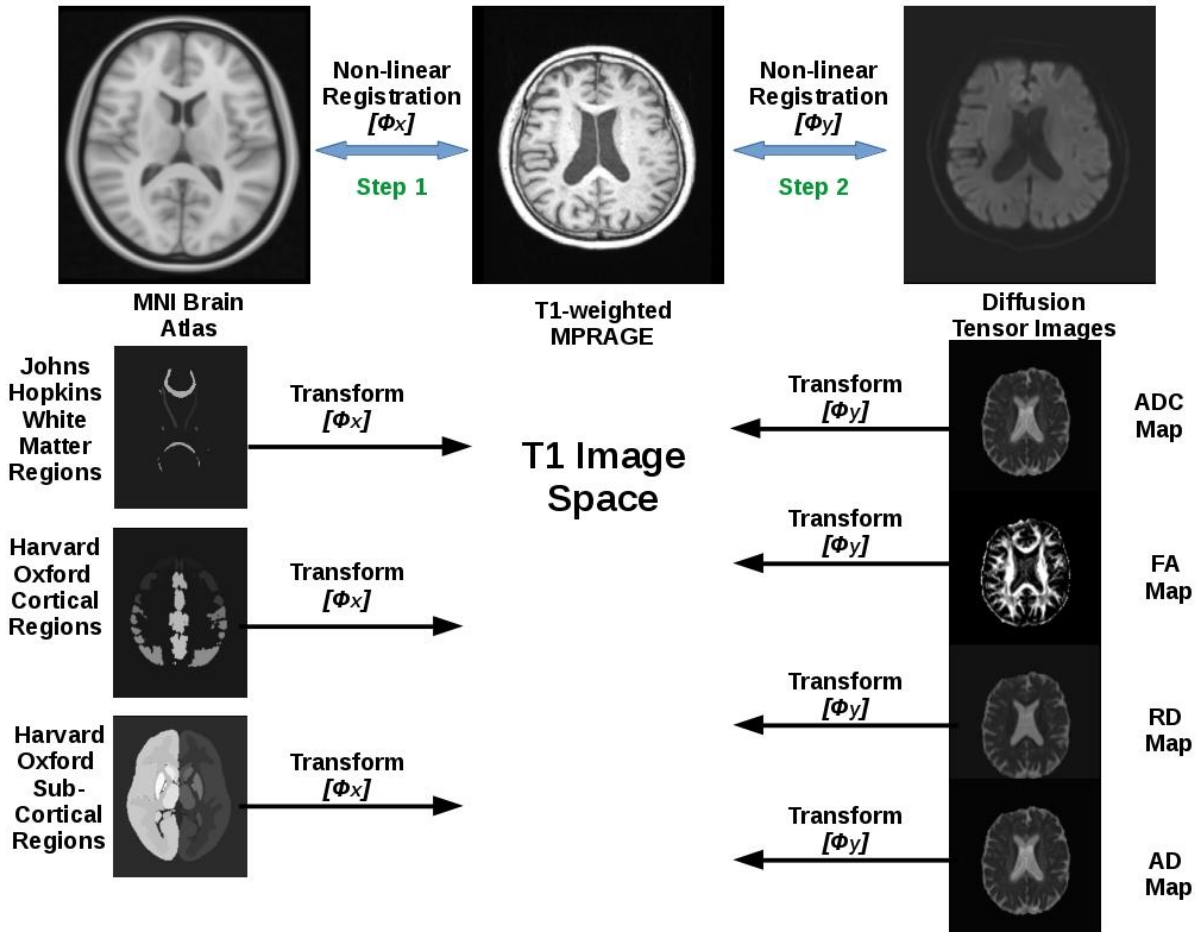

FIGURE 3.2. NON-LINEAR REGISTRATION OF MNI ATLAS TO T1-WEIGHTED MR IMAGE USING ANTS AS WELL AS THE NON-LINEAR REGISTRATION OF VARIOUS DIFFUSION MAPS TO THE NATIVE T1-WEIGHTED PATIENT SPACE

The Harvard-Oxford cortical, Harvard-Oxford sub cortical, and the Johns Hopkins University (JHU) white matter tractography atlas brain regions were transformed into the T1 space using the transformation obtained from the MNI to T1 registration. The JHU atlas includes 20 additional

structural tract information in regions such as anterior thalamic radiation, forceps major/minor, superior longitudinal fasciculus and others and have been widely used in DTI based studies.

Similarly, each patient-specific DTI dataset was registered to the corresponding MPRAGE dataset using another non-linear registration. This non-linear registration as well as eddy current distortion correction was used to compensate for inherent DTI-B0 distortion effects. More precisely, the average ($b=1000s/mm^2$) DTI image was used as the reference for this due to improved anatomical details and higher similarity to the T1-weighted MPRAGE dataset. This registration also consisted of a rigid followed by an affine transformation used for initialization of the non-linear registration, which was performed using a symmetric diffeomorphic image registration method.[114]

The diffusion parameter maps were transformed to the MPRAGE dataset using the corresponding non-linear transformation. The transformed brain atlas regions from the first registration step were then used to determine median diffusion parameters such that 516 DTI features are available for each patient. Median instead of average values were used to account for potential non-normal DTI parameter value distribution and partial volume effects at the border of brain structures. Similarly, the registration and analysis process of the R2 image datasets was identical to the one described in the DTI registration, with the exception that the Johns Hopkins University white matter tractography atlas brain regions were excluded. As mentioned before, the R2*, R2prime, R2' maps were calculated from the raw T2-weighted images using the ANTONIA software developed by Forkert et al.[110]

4.2.3 Feature Extraction

An overview of the extracted features using different atlases are presented in Table 3.2. In order to extract the feature set from the T1- and T2-weighted and DTI images, a custom, in-house developed Python script was used. In detail, the code reads the registered atlas regions (which are different depending on each feature type) and superimposes them onto the target quantitative maps (i.e. Volumetric, ADC, FA, R2*, etc.). After this stage, as each region of interest on the atlas has a specific intensity value, the atlas is thresholded based on that value and the resulting binary (where voxel have values 1 if they belong to the specific region of interest and 0 if they are not) mask is then multiplied with the quantitative map. Consequently, for calculation of brain iron content markers as well as diffusion parameters, each voxel value in that specific ROI is stored in an array and the median value is calculated and stored in a separate array. For the extraction of morphology-based features, a similar approach was taken. However, in this case, the number of voxels within each ROI was counted and multiplied by the constant voxel volume. The voxel volume was calculated as the multiplication of the native X, Y, and Z resolutions. Consequently, a measure of volume was acquired for each of the analyzed ROIs. In addition to the parcellated atlas brain regions, the non-linear deformation field was also used for transforming a binary segmentation of the total intracranial volume to each patient dataset, which was used for extended morphometric analysis as well as volumetric normalization of the single brain regions to account for differences regarding the general head anatomy. Apart from the volume, the segmented brain regions were also used to determine the surface area of each brain region as well as the surface-area-to-volume ratio (SAV). For calculation of the SAV, the raw regional volumes instead of the volumes corrected for the full intracranial volume were used.

Unlike the previously extracted regions, volumetric features were not separated in the right and left hemispheres due to fact that most previous studies have focused on entire regions. The Johns Hopkins white tractography atlas was not used here as well because the segmented regions are not large enough and therefore have little volumetric utility. Finally, all the results were stored in a separate text file and rearranged in a comma separated format (CSV). The CSV format was chosen due to its simplicity as well as readability in the classification framework explained below. As it was mentioned before, different atlas segmentations were used to extract the corresponding feature maps. In detail, for morphometric features, the Harvard-Oxford cortical and sub cortical atlases were used. Unlike the iron markers and diffusion features, the Johns Hopkins white tractography atlas was not used due to the small tract-based segmentations it provides, which have no morphometric information. Similarly, for the iron marker and diffusion features, the Harvard-Oxford cortical, sub cortical, and Johns Hopkins white tractography were employed. However, the Harvard-Oxford cortical atlas was separated into left and right regions. The rationale behind this approach is that previous studies have highlighted group-wise differences between PD, HC, and PSP-RS in left and right regions in terms of iron marker and diffusion.[35,115,116] In the diffusion based features, the ventricle were excluded as they contain no meaningful information.

Table 3.2. In-depth information regarding the features used in this study

Modality	Type of Feature	Atlases Used	Number of Features per category	Overall Category Features
T1-Weighted	**Morphometry:** Regional Volume (V) Regional Brain Surface (S) Regional Surface area to Volume Ratio (SaV)	Harvard-Oxford Cortical	48	234 (V, S, SA:V)
		Harvard-Oxford Sub Cortical	21	
		MNI brain regions	9	
			Total: 78	
T2-Weighted	**Median Regional Brain Iron markers:** R2, R2*, R2'	Harvard-Oxford Cortical	94 (47 *2/left and right)	396 (R2, R2*,R2')
		Harvard-Oxford Sub Cortical	21	
		Johns Hopkins University White Matter Tractography	17	
			Total: 132	
DWI	**Median Regional Brain Diffusion Levels:** MD, FA, AD, RD	Harvard-Oxford Cortical	92 (46 *2/left and right)	520 (MD, FA, AD, RD)
		Harvard-Oxford Sub Cortical	21	
		Johns Hopkins University White Matter Tractography	17	
			Total: 130	

4.2.4 Classification Pipeline

Following the feature extraction stage explained in section 3.2.3, each of the HC, PD, and PSP-RS subjects has a certain number of features per category. The morphological analysis results in a total of 234 imaging features consisting of 78 features for each of the volume, brain surface area, and SA:V categories. In detail, 48 cortical, 21 sub-cortical structures as well as 9 MNI brain regions were determined per patient in each category. In addition, a total of 396 T2-weighted features including 132 R2, 132 R2', and 132 R2* features, were included. An overall number of 520 DTI features including cortical (n=92), sub-cortical DTI (n=21), and n=17 John Hopkins University white matter tract features are available per subject, consisting of 130 ADC, 130 FA, 130 AD, and 130 RD features.

Feature selection block

In this stage, all the 1150 features in each category are passed through a feature selection block. In detail, features are ranked based on their relevance with respect to the following feature selection algorithms: correlation attribute evaluator, gain ratio/information gain evaluator, principle component analysis, RELIEFF, and support vector machine attribute evaluator. The underlying basic concepts of each of these methods were described previously in section 2.4.2. Following each of these 6 feature selection methods, the features are ranked based on their importance in terms of class differentiation. A ranker essentially sorts features based on their classification relevance according to the aforementioned feature selection methods. Here, the top ~10% ranked features (100 features in total) were selected for initial inclusion in the

classification. This number was selected in order to decrease computational processing times associated with higher number of features.

Classification/training block

Each ranked feature set is then used to train each of the following classification models: decision tree, random forest, logistic model tree, k-nearest neighbors, naive Bayes, support vector machine, and multi-layer perceptrons. The leave-one-out cross validation routine was employed for classifier performance evaluation in both the training and testing phase. Furthermore, to prevent double dipping, the leave-one-out cross validation (LOOCV) also included the feature ranking as described above so that the optimal features used for the actual classification can vary in each iteration of the leave-one-out cross validation. The optimal number of highest ranked features used for training and testing of the classifier was systematically optimized by iteratively removing the lowest ranked feature from the training and testing. This recursive feature elimination approach was performed by continuously removing the features from the initial top 100 down to the top 10 features in each category. Ultimately, a custom code written in python was utilized to detect the "feature selection-method + classification method" pair with the best classification performance. The overall 3-level classification accuracy was selected as the metric evaluating classification performance. In case of equal classification performance, the feature selection + classification pair with the lowest number of features were reported following the Occam's razor principle. This principle states that the more assumptions (in this case features) are made, the more unlikely an explanation (classification) becomes. Once the top performing model is identified, a 1000 permutation test was applied to test for statistical significance of the

classification. In detail, the original class labels in the initial dataset were randomly shuffled so that 1000 random variations of the initial datasets were available. Consequently, the top performing model was tested with the aforementioned random datasets and the obtained accuracies were recorded. We say the classification is significant (and therefore not a result of noise) if the model performs in the top 5% of all the 1000 results.

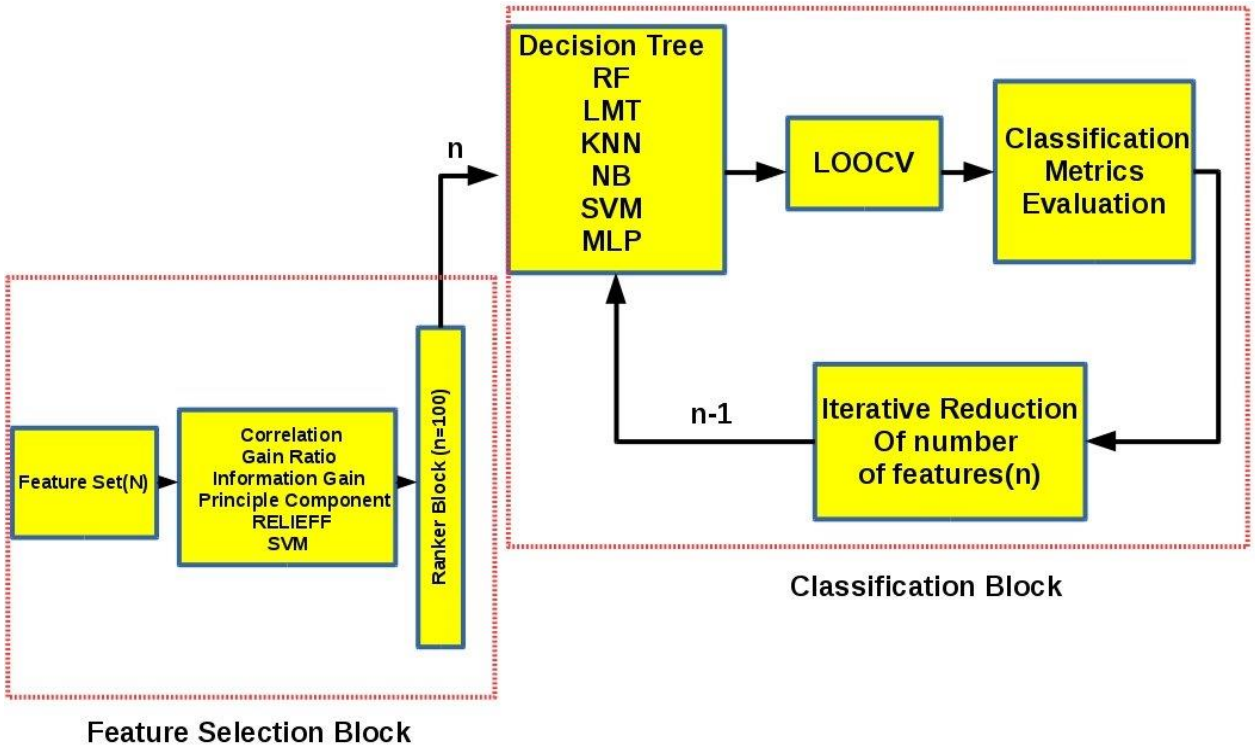

FIGURE 3.3. CLASSIFICATION PIPELINE CONSISTING OF FEATURE SELECTION AND CLASSIFICATION BLOCKS. THE TOP PERFORMING FEATURES ARE ASSESSED VIA THE RECURSIVE FEATURE ELIMINATION APPROACH

CHAPTER FIVE: RESULTS AND ANALYSIS

5.1 Classification Results using Single Modalities

In the first stage, the classification pipeline described in section 4.2.4 was utilized to perform classification between HC, PD, and PSP-RS only using single modality features (T1-weighted, T2-weighted, DTI). In the following sections, the best performing feature selection and classification pair in each instance, the top performing features, the overall observed trends, and the corresponding classification metrics are described.

T1-weighted MRI

Cortical, sub-cortical, and MNI based regional brain volume, surface area, and surface area to volume ratio features derived from structural T1-weighted images resulted in a top overall accuracy of 65.04% between the three classes. The contingency matrix of this classification task is presented in table 3.3. This accuracy was achieved when the top 40 features were selected following a gain ratio feature ranking and the SVM classification approach. In detail, the differentiation between healthy and different diseased states was rather poor, amounting to a total of 15 misclassified healthy controls out of 38. Furthermore, classification of PD was also sub-optimal, resulting in 10 and 7 PD subjects being wrongly classified as HC and PSP-RS, respectively. The top ranked features consisted mainly of sub-cortical structures such as brainstem, pallidum, putamen, thalamus, and cortical structures, including occipital pole, frontal and temporal gyrus, opercular gyrus, and others. In addition, volume, surface area, and SA:V features were equally present in the ranking, signifying the importance of including various multi-aspect (not multi-source) T1-weighted features.

Table 3.3. Confusion matrix following a gain ratio + SVM classification combination using morphological features only

Morphology Features (Surface area, Volume, and Surface-Area-to-Volume Ratio Features)											
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	HC	PD	PSP	Accuracy
HC	0.605	0.185	0.657	0.605	0.630	0.429	0.710	23	12	3	65.04%
PD	0.622	0.241	0.667	0.622	0.644	0.384	0.690	10	28	7	
PSP	0.800	0.120	0.615	0.800	0.696	0.619	0.840	2	2	16	

T2-weighted MRI

The classification metrics based on the R2, R2' and R2* maps are described in Table 3.4. Following this process, the highest performing classification accuracy was obtained when the top 30 features were used. The principle component feature selection method in combination with an LMT classification resulted in the highest performance of 75.72%. In detail, 9 healthy controls were misclassified as diseases (7 PD and 2 PSP-RS), whereas 7 and 2 PD subjects were wrongly classified as HC and PSP-RS, respectively. In addition, utilizing T2-weighted features resulted in 5 PSP-RS patients being wrongly categorized as PD. Overall, weighted combinations of brain iron measures in the cerebral cortex, left accumbens, left amygdala, and left hippocampus were present in the highest ranked PC features.

Table 3.4. Confusion matrix following a PCA + LMT classification combination using brain iron content measures only

T2-weighted Image Features (based on Quantitative R2, R2', and R2* Features)											
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	HC	PD	PSP	Accuracy
HC	0.763	0.108	0.806	0.763	0.784	0.663	0.875	29	7	2	75.72%
PD	0.756	0.207	0.739	0.756	0.747	0.547	0.845	7	34	4	
PSP	0.750	0.072	0.714	0.750	0.732	0.665	0.948	0	5	15	

Diffusion-Tensor Imaging (DTI)

The classification results based on the DTI measurements (ADC, FA, RD, AD maps) are shown in Table 3.5. The highest performing classification accuracy was obtained when the top 69 features were used. In detail, the information gain-based feature ranking method in combination with a LMT classification resulted in the highest performance of 95%. Overall, a total of 5 subjects were misclassified (3 PD and 2 PSP-RS), whereas no healthy controls were placed in the diseases categories. The highest-ranking diffusion features included mostly cortical regions such as the parahipocampal gyrus, cingulum, cingulate gyrus, opercular cortex and others as well as sub-cortical structures such as the thalamus, brainstem, and pallidum.

Table 3.5. Confusion matrix following an information gain + LMT classification combination using diffusion MRI maps only

Diffusion Tensor Imaging Features (ADC, FA, RD, AD Features)											
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	HC	PD	PSP	Accuracy
HC	1.000	0.000	1.000	1.000	1.000	1.000	1.000	38	0	0	95.14%
PD	0.933	0.034	0.955	0.933	0.944	0.901	0.975	0	42	3	
PSP	0.900	0.036	0.857	0.900	0.878	0.848	0.968	0	2	18	

5.2 Classification Results by Combining Multiple Modalities

The classification results based on the complete feature set, utilizing information from T1-weighted MRI (brain surface area, brain volume, surface area to volume ratio), T2-weighted MRI (brain iron content measures) as well as DTI (ADC, FA, RD, AD maps) are shown in Table 3.6. The highest classification accuracy in this multi-source model was obtained when the top 79 features were used. In detail, the support vector machine feature ranking method in combination with a MLP classification resulted in the highest performance of 95.14%. Overall, a total of 5 subjects were misclassified including 1 PD and 4 PSP-RS subjects, whereas none of the healthy controls were placed in the diseases categories. In detail, 11 attributes (~13%) from the total set of 79 were related to morphological features, including deep grey matter regions such as left pallidum, left and right thalamus, left caudate and brainstem. Several cortical structures such as precentral and supramarginal gyrus, angular gyrus, temporal fusiform cortex, planum polare were also among the top ranked features. Moreover, most of the morphological features selected

for this classification problem included volume values rather than brain surface area or SA:V. In terms of brain iron accumulation markers, a total of 12 features amounting to ~15% were present in the optimal feature set. Brain iron content markers in areas such as the temporal fusiform, parahippocampal gyrus, frontal gyrus and others were ranked high according to SVM based feature selector. The rest of the top ranked features were diffusion features. Diffusion attributes in regions such as parahippocampal gyrus, insular cortex, pallidum, thalamus, brainstem, putamen, and others were ranked as the most discriminative features. Table 3.7 represents an overview of the feature composition in the top performing feature selection and classification pipeline.

Table 3.6. Confusion matrix following an SVM based feature selection + MLP classification combination using features from multiple MRI modalities

Combination of All Imaging Features (Morphology, Brain Iron Content Marker, Diffusion)											
Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	HC	PD	PSP	Accuracy
HC	1.000	0.000	1.000	1.000	1.000	1.000	1.000	38	0	0	95.14%
PD	0.978	0.069	0.917	0.978	0.946	0.904	0.986	0	44	1	
PSP	0.800	0.012	0.941	0.800	0.865	0.840	0.983	0	4	16	

Table 3.7. Feature composition of the SVM based feature selection + MLP classification combination using features from multiple MRI modalities

Modality	V	S	SaV	R2	R2*	R2'	FA	MD	AD	RD	Total
T1-weighted	6	2	3	-	-	-	-	-	-	-	11
T2-weighted	-	-	-	2	3	7	-	-	-	-	12
DTI	-	-	-	-	-	-	15	12	13	16	56

CHAPTER SIX: DISCUSSION AND CONCLUSIONS

6.1 Discussion

Inadequate levels of accuracy in PD subtype diagnosis through traditional criteria such as the UK Parkinson's Disease Society Brain Bank, Unified Parkinson's Disease Rating Scale, and others in combination with standard anatomical MRI have promoted the use of PD-CAD systems. In this context, neuroimaging data obtained from various MRI modalities such as T1-, T2-, and diffusion-weighted MRI are among the most useful to diagnose PD. Moreover, biomarkers extracted from other imaging protocols such as susceptibility-weighted imaging (SWI), quantitative susceptibility mapping (QSM), and functional MRI (fMRI) have also been proposed as viable image-based methods to diagnose and differentiate PD syndromes.[87,105,117,118]

In the current research project, a three-class individual-level classification between PD, PSP-RS, and healthy controls was developed and evaluated using multi-modal features obtained from T1-, T2-, and diffusion-weighted MRI. Following the extraction of these features, a wide variety of feature selection and classification methods were employed to identify the top performing combination in terms of differentiation accuracy. The aforementioned pipeline was repeated four times with different input features. In the first stage, only morphological features derived from structural T1-weighted MRI datasets including brain volume, brain surface area, and brain surface area to volume ratio were utilized for training and testing of the classifiers. Consequently, following a grain ratio feature ranking and SVM classification combination, a top overall accuracy of 65% was attained whereas the differentiation between PD and HC was especially poor, effectively downgrading the overall classification accuracy. In detail, only 23

out of 38 healthy controls were correctly classified. The PD vs. PSP-RS differentiation was also weak, amounting to a total of 16 misclassified cases out of the 65 diseased (PD + PSP-RS) cases. In the second stage, brain iron content markers from the T2-weighted MRI protocol were employed for classification. Using these features, a higher differentiation accuracy of 75.72% was achieved using a combination of PCA feature selection method and the LMT classifier. Similar to the morphology-based classifier, this rather low accuracy can be mainly attributed to the weak differentiation between PD and HC as well as PD and PSP-RS. While the classification accuracy between PD and HC improved compared to the morphology-based classification, 7 cases in each category were still misclassified as PD and HC (14 in total). The PD vs PSP-RS differentiation accuracy remained the same as the morphology-based classification with a total of 9 misclassified instances. In the third stage, diffusion-based parameter maps such as FA, MD, AD, and RD were used as features for the classification. The combination of information gain feature ranking and LMT classifier resulted in a top accuracy of 95.14%, effectively outperforming the previous two classification setups. Only five non-healthy patients were misclassified whereas 3 PD and 2 PSP-RS subjects were incorrectly classified as PSP-RS and PD, respectively. Ultimately in the last step, all features from the previous three setups (i.e. morphology, brain iron content markers, and tissue diffusion properties) were combined and implemented within the pipeline. Using this comprehensive feature set, an accuracy of 95.14% following an SVM feature ranking and the MLP classification method was achieved. Overall, 1 PD and 4 PSP-RS subjects were incorrectly classified as PSP-RS and PD, respectively. The obtained accuracy of this multi-modal classifier was identical to the performance achieved when only DTI features were employed. However, the disease categorization differed between the two

cases. Consequently, the combination of multi-modal features outperforms the single modality features in terms of PSP-RS specificity. In the following sections, the implication and possible interpretations of the present study will be discussed.

T1-weighted Features

The highest accuracy in PD vs. HC using T1-weighted based features was obtained by Singh and Samavedham[119] at 99.78%. The presented results in this study, however, are strongly suspected of a potential over-fitting case as the least square SVM kernel employed here is known to be prone to over-fitting and therefore the classification is likely to be unreliable if a completely separate testing dataset is employed.[120] Apart from the aforementioned study, several other individual level classification studies using volumetric features in PD vs. HC have reported accuracies ranging from less than 50% up to 86%.[75,76,81,83,121] As a general pattern however, these volumetric-based studies were unable to achieve reasonable classification accuracies differentiating PD vs. HC only using volume-based features. In fact, the results obtained in this study in terms of PD vs. HC differentiation using morphology features, seems consistent with previous studies. The proposed morphology method performed poorly resulting in a total of 22 misclassified cases from a total of 83 subjects (~26% error). One potential reason for the poor differentiation ability of morphology measure in PD vs. HC could be related to less pronounced structural differences in PD compared to HC.[122,123] Several studies have shown that noticeable structural changes in PD do not manifest until later in the disease course. Consequently, macro-structural features (i.e. morphology) are not sensitive enough to differentiate PD from HC. In terms of PD vs. PSP classification, individual level classification methods using morphology

features as previously described by Scherfler et al.[73], Sarica et al.[75], and Focke et al.[25] were able to obtain comparably high classification accuracies. Unlike PD vs. HC, it appears that due to the clear structural differences between PD and PSP-RS, features derived from high resolution T1-weighted images are indeed valuable for differentiation between these two syndromes. Similarly, the proposed model in this research project performed relatively well, only misclassifying 13% of the total patients. Finally, studies focusing on PSP vs. HC classification have resulted in even higher accuracies of up to 94%.[75,76,82] In detail, the highest accuracy for classification of PSP vs. HC with 93.75% was previously achieved by Sarica et al.[75] by using volumetric features in a naive Bayes classifier validated by a ten-fold cross validation in a relatively large cohort of 46 HC, 65 PD, and 32 PSP. These results are to be expected considering the significant structural and functional differences in PSP compared to HC.[124,125] In the present study, a classification accuracy of 89% was attained for PSP-RS vs. HC classification in a three-level classification routine, also including healthy controls. Due to the clear potential for differentiating PSP-RS subjects from healthy controls and PD subjects, incorporations of morphological features within a PD-CAD framework appears still viable and can add to the comprehensiveness of such devices.

In this research project, the differentiation abilities of a wide range of cortical and sub-cortical volumetric features were investigated. In fact, this is also one of first studies to investigate the potential benefits of using other less frequently used morphological parameters such as brain surface area and surface-area-to-volume ratio for classification. Previous studies focused mostly on the volume of specific regions of interests, such as the brainstem, which are known to be significant discriminators between PD and PSP.[28–30] However, recent studies suggest that various

other brain regions, apart from the deep gray matter, are also affected by atrophy and could hold additional informative value as input features in classification routines.[23,27,34] In addition to plain volume of brain structures, previous research suggests that the corresponding surface area of the brain structures might also have some additional predictive value for the classification of neurological diseases.[116–119] Worker et al.[26] for example, found an increased surface area in the pericalcarine cortex in PSP compared to PD patients along with general patterns of cortical thinning as well as volume loss in the superior gyrus in PSP compared to PD. Nevertheless, the surface area features and combination of volume and surface area in terms of the surface-area-to-volume ratio (Sa:V) have only been barely used for classification of PD syndromes. The main findings of this research project with implications to future developments of morphological classification methods for HC vs. PD vs. PSP-RS differentiation are four-fold. First, the inclusion of the surface area and SA:V values provide complimentary information and leads to better classification results compared to using volumetric features only. This can be inferred by investigating the top 40 selected features that resulted in the best classification accuracy. The top features consist (almost evenly) of a wide variety of volumetric, regional brain surface area, and SA:V features. This finding highlights the benefits of including multi-faceted features that can be obtained from standard T1-weighted images. Second, adding the morphological profiles of cortical structures for classification does not seem to improve the classification accuracy compared to using all three sub-cortical morphological features and can potentially be neglected. Looking back at the top 40 features, it is evident that most selected features belong to sub-cortical rather than cortical regions. Another way of interpreting these results is that sub-cortical regions are more macro-structurally impacted by the Parkinsonian syndrome than cortical

structures, thus medical interventions and future studies could mainly focus on these regions. Third, the employed gain ratio feature ranking method was able to select the most well-known PD regions in a purely automatic fashion. Many studies in the past have resorted to investigating certain ROIs of disease specific regions that were found to be important according to group-wise studies and have selected those as input features for classification.[25,73,75,82] Clearly, the presented method is much more objective than the manual approach as it enables the investigation of a wide set of anatomical features in little time without a selection bias. The highest ranked morphology parameters are brainstem, pallidum, thalamus, putamen, and hippocampus, all of which are known to be impacted in PD syndromes.[126] Finally, as indicated by table 3.3, it is clear that morphologic measures are indeed not beneficial for the classification of HC and PD. Nearly half of the HC subjects were misclassified as PD, which ultimately resulted in the poor overall three level classification performance using features obtained from this modality. Consequently, the usage of volumetric features for the classification of PD vs. HC is not recommended. Meanwhile, the differentiation between PD vs. PSP-RS vs. HC was much better and consistent with existing literature. It seems that features obtained from other MRI modalities are more useful in the classification of PD vs. HC.

T2-weighted Features

In a similar study to this research, Boelmans et al.[85] utilized brain iron content markers in a logistic discriminant analysis approach to classify 24 HC, 30 PD, and 12 PSP subjects. Subsequently, the classification resulted in an overall accuracy of 74.2%. Moreover, using a similar classification pipeline, Eckert et al.[86] achieved a total accuracy of 75.4%. However, the

latter study consisted of a different sample size of 20 HC, 15 PD, 10 PSP, and 12 MSA. None of the two studies mentioned utilized a feature section method but rather opted for a region of interest approach. Studies following a region of interest approach usually investigate a number of well-known PD-impacted regions and neglect other less known regions. In this research thesis, a similar accuracy of 75.72% was obtained using a sample size of 45 PD, 38 HC, and 20 PSP-RS following a PCA feature selection and LMT classification method. While the accuracies obtained are similar, there are two significant differences between those studies and this research project, ultimately elevating the current method over previously reported results. First, the patient sample size of this study is comparatively larger than previous research, effectively lending the model a higher level of generalizability. Further re-enforcing the generalizability of the present method, is the fact that none of the previous studies performed cross validation principles, whereas leave-one-out-cross-validation was used in this study. Second, input features in the two previous studies were handpicked and not automatically identified as done in this work. Consequently, the current method will be preferred over previous ones due to its reliability and comprehensiveness. Moreover, the feature selection method analyzes the iron content of a wide spectrum of brain regions, thus providing a broader overview of potentially impacted regions.

In terms of PD vs. HC classification, the proposed model performs rather poorly with 7 PD subjects wrongly classified as HC and vice versa. The misclassification in this category was mostly responsible for downgrading the overall classification performance. A potential explanation for this result could be that iron content markers were not pronounced enough in the PD group, therefore resulting in sub-optimal classification performance. Similar to morphological information, brain iron markers seem to be inadequate features to differentiate

between PD and HC. The differentiation of PSP-RS vs. PD was rather modest with 4 PD and 5 PSP-RS subjects wrongly classified as PSP-RS and PD, respectively. This is an interesting finding considering that morphological features also resulted in a total of 9 misclassified instances differentiating PD vs. PSP-RS. However, the difference is that misclassified patients using brain iron marker features were rather balanced between the two diseases, whereas morphological features exhibited a weaker PSP-RS differentiation. More research needs to be focused on investigating the rate of atrophy (i.e. morphological cell loss) compared to the accumulation of iron markers as this will clarify some of the ambiguities surrounding this finding. Ultimately, these results seem to once again encourage the combination of morphological and brain iron markers for an improved PD vs. PSP-RS differentiation.

In addition, the differentiation of HC vs PSP-RS was far better resulting in only 2 PSP-RS subjects misclassified as HC. It could be inferred that iron content markers are significantly different in the PSP-RS group compared to HC, which is consistent with current literature [36] or that certain regions in PSP-RS have a significant iron accumulation not present in the HC group. The obtained classification accuracy here is similar to the accuracy obtained using the morphological features. It seems that structural and brain iron markers are significantly different between the two groups, effectively resulting in high levels of classification accuracies. All in all, brain iron content markers by themselves are not reliable features for the differentiation of HC from PD and PSP-RS from PD. Consequently, it seems that the addition of features from other sources might hold complimentary information, effectively improving individual level differentiation between the groups. A limitation of the presented pipeline however, is that the feature selection method used is the principle component analysis. As previously mentioned,

following PCA, new features are created from a linear combination of the original dataset. Consequently, it is not clear how exactly individual features from the initial feature set contribute to the overall classification.

Diffusion Features

Previous studies have shown the benefits of using diffusion-based features for the classification of PD and healthy controls. Scherfler et al.[89], Salamanca et al.[90], and Banerjee et al.[93] used diffusion measures such as MD (or ADC) and FA parameter maps for classification and obtained accuracies of up to 98%. In the present study, the top accuracy of 100% was attained for the classification of PD and HC using an information gain feature ranking and LMT classification approach with a sample size of 45 PD and 38 HC. This accuracy was obtained when features from ADC, FA, RD, and AD maps were incorporated within the classification framework. The cohort size used in this research project was larger than the studies of Scherfler et al. and Banerjee et al., meaning that the proposed pipeline has higher generalizability potential. In addition, the obtained results for PD vs. HC differentiation is even higher than the multi-source approach taken by Peran et al.[96]. In the study by Peran et al., features from T1-weighted, T2-weighted and diffusion-weighted MRI were employed for the classification of 22 HC and 30 PD subjects following a multi-parametric regression analysis, resulting in an accuracy of 95%. It seems obvious that the differentiation potential of DTI measurements is much higher than the previously discussed morphological and brain iron content marker features. One potential reason for this finding is that micro-structural changes occur earlier than macro-structural changes. T1- and T2-weighted MRI sequences are modalities for investigating macro-structural changes,

therefore they are less informative when the classification task contains classes with similar profiles. Consequently, as DTI offers measures for tissue integrity at a micro-structural level, it appears to be able to perform PD vs. HC classification at higher accuracies compared to the other feature groups (volumetric from T1-weighted or iron content markers from T2-weighted)

With respect to differentiating PD and PSP-RS syndromes, 2 PSP-RS and 3 PD subjects were wrongly classified as PD and PSP-RS, respectively. The obtained sub-syndrome classification accuracy is among the top reported results thus far. It should be noted that this is not the first work to employ DTI measurements for classification of PD and PSP-RS subjects. Haller et al.[87] presented an approach to classify PD subjects (n=17) and subjects with atypical forms (n=23) of Parkinsonism using the RELIEFF feature selection method, a support vector machine classifier, and voxel-wise FA values as features. A correct classification between PD subjects versus subjects with atypical forms was achieved at up to 97.5±7.5%. However, it should be noted that the group of 23 subjects with atypical forms of Parkinsonism included only one patient with PSP while the other subjects in this group were, for example, diagnosed with multiple system atrophy, dementia with Lewy bodies, vascular Parkinsonism, and even traumatic brain injury. Thus, the results are not comparable to those described in this work. Furthermore, using voxel-wise features for classification always bares the risk of over-fitting.[107] Similarly, a 100% accuracy in the classification of PSP vs. HC was achieved in this research project. To the best of my knowledge, no classification task has directly used diffusion metrics to classify these two groups. Moreover, due to the distinct micro- and macro-structural differences between the two groups, traditional clinical diagnosis methods also perform well in this category.

After investigating the results of the gain feature ranking and selection method, several patterns can be observed. Looking back at the results obtained by the classification using morphological features, it was shown that mostly sub-cortical features were discriminative. Conversely, it appears that diffusion changes are in fact a global effect as cortical as well as sub-cortical structures brain regions were ranked highly based on the information gain method. As micro-structural changes are typically expected to occur prior to measurable macro-structural changes, DTI parameters might be more viable as early disease biomarkers to differentiate between PD and PSP-RS. The 69 brain regions selected and used for classification included the brainstem, deep gray matter structures such as thalamus, putamen, and pallidum all of which are known to be affected by PD and have previously been identified as important brain regions for the volumetric differentiation of PD vs. PSP.[25,27–30,127] This might further corroborate the proposition that micro-structural changes manifest earlier than or are at least correlated with macro-structural changes, promoting the use of diffusion-based sequences over the traditional morphology or brain iron content marker features for differential diagnosis of PD vs. PSP-RS. However, this speculation needs to be investigated in more detail in future studies. The other brain regions identified by the feature selection method are part of the frontal cortex, namely the superior frontal gyrus and frontal medial cortex, which could be related to previously reported differences in the prefrontal dopaminergic system between PD and PSP subjects.[128] It should also be highlighted that this is one of the first studies to investigate the potential benefits of using the full range of DTI parameter maps (FA, MD, AD, RD) in conjunction with a feature selection method. Most previous studies have only used FA and MD values for classification or group-wise studies. However, as previously mentioned, the incorporation of multi-facet features within

a given modality (as it was the case with the morphological features in this study) has the potential to improve the comprehensiveness of a machine learning model as it helps elucidating several aspects of the PD syndrome.

Multi-Modal Features

In terms of classifying PD vs. HC by combining image features from multiple sources, several studies have been performed in the past. Long et al.[105], utilized structural and functional MRI information to classify PD and HC subjects. The results reported show that the combination of features resulted in an improved classification of 86.92%. In a study similar to this research endeavor, Peran et al.[96], combined T1- weighted, T2-weighted, and diffusion-weighted MRI features to obtain an PD vs. HC classification accuracy of 95%. However, the HC vs. PD classification task in this study is superior to previous studies due to the obtained 100% accuracy by using a leave-one-out-cross validation in a much larger sample size. Furthermore, unlike the aforementioned studies that performed binary classifications, the differentiation task in this study was inherently more complex as three classes (HC, PD, PSP-RS) were investigated. Moreover, in the differentiation of HC vs. PD, the combinational feature method in this research thesis outperformed single modality approaches using morphological and iron content markers alone and ultimately tied with diffusion modality features only.

In case of PSP vs. PD, multiple studies have also attempted a multi-modal approach. Morisi et al.[97], combined volumetric, diffusion, and proton spectroscopy measures in a linear kernel SVM and obtained an accuracy of 98%. Cherubini et al.[101], used volumetric and diffusion features from a large sample of 57 PD and 21 PSP and achieved an accuracy of 100% using only white

matter features. In the present study, a PSP vs. PD differentiation of 93% was achieved. The obtained results in this study outperforms the results of Morisi et al. due to the fact that a larger sample size and a more rigorous validation approach was used. The results of Cherubini et al. are somewhat peculiar as the presented analysis is vague in parts and do not fully clarify the entire classification procedure. The important aspect of the presented results is that the multi-model approach does not outperform the single modality diffusion method as they both are tied as the top performing classification model overall. However, the PD and PSP-RS differentiation profile differs between the two models. In case of using diffusion features only, 3 PD and 2 PSP-RS subjects were wrongly classified as PSP-RS and PD, respectively. If we consider the number of correctly classified PD and PSP-RS subjects as true positive (TP) and true negatives (TN), respectively we can calculate sensitivity and specificity as well as a weighted average between them. Consequently, following this definition, the sensitivity and specificity are 95.45% and 85.71%, respectively. Alternatively, if the entire feature set is employed the 4 PSP-RS and 1 PD subjects are incorrectly classified as PD and PSP-RS. This translated to a sensitivity and specificity of 91.67% and 94.12%, respectively. Based on the obtained results in this study, the inclusion of multi-model feature sets has a major practical advantage over the use of single modality features. It seems that the combination of multiple features from various sources increases PSP-RS specificity by about 10% compared to using only diffusion-based features while the sensitivity is decreased by only 4%. This is an interesting finding considering that PD sensitivity is already quite high based on established clinical criteria, whereas specificity is rather low. For instance, the so-called “bicycle riding test” has a PD sensitivity and specificity of 96% and 52%, respectively.[129] Another example, is the levodopa responsiveness, which has a PD

sensitivity of 94%, but a very low specificity of only 30%.[130] A key point regarding levodopa responsiveness is that even though 94% of true PD patients show a positive response to levodopa, other parkinsonian sub-syndromes such as multiple system atrophy and dementia with Lewy bodies might also have a positive "initial" response but ultimately become unresponsive to levodopa as the disease progresses. Conversely, even if we flip the definition of true positive and true negative so that this time the number of correctly classified PSP-RS and PD subjects are true positive (TP) and true negatives (TN), respectively, we will still observe that the multi-modal approach results in higher overall sensitivity and specificity, which outperforms the DTI based classification and the previously aforementioned bicycle and levodopa tests. Moreover, according to table 3.7, following the multi-modal approach, 11 morphometric, 12 iron content marker, and 56 diffusion features were included in the top performing machine learning pipeline. The fact that all three feature categories have a presence in the top performing features (some more than others) clearly indicates the benefits of including multi-modality features. Not surprisingly, the number of DTI features, which are the majority in the multi-modal approach, is sensible considering that classification using only DTI features has a high performance compared to the morphometric and iron marker-based classification.

Therefore, considering that the aim of this research thesis is to present a framework of "early" PD-syndrome detection, it seems highly likely that future iterations of this study will result in higher and more robust differentiation performance. Consequently, the improved specificity is highly beneficial to clinical assisted diagnosis procedures. All in all, while the obtained accuracy is identical for the two top performing cases (DTI only and multi-modal), the inclusion of multi-modal features improves specificity.

Feature Selection and Multi-Level Classification

As an integral part of many classification techniques, several feature selection methods such as information gain, principal component analysis, linear SVM based feature selection, RELIEFF, Fisher vector algorithm, evolutionary based techniques, fuzzy based data transformation, graph theory methods, and others were previously employed in previous studies.[63,77,131] However, many previous studies did not incorporate any feature selection algorithm and employ pre-selected sub-syndrome specific features from a limited number of ROIs instead. There are two main advantages to feature selection steps within CAD-PD systems. First, feature selection methods are capable of analyzing information in large scales and are able to find patterns that might not be evident to clinicians. Therefore, the selection of only a few ROIs will reduce the analysis to those regions only and consequently other less frequently observed regions will be left out, thus potentially limiting the comprehensiveness of a CAD-PD device. Therefore, the acquisition of a broad range of features that can be used to quantify various PD subtypes is recommended. In the next step, feature reduction can be implemented to reduce data dimensionality. In the present study, multiple feature selection methods were used to identify the most discriminative features. In fact, most of the top selected features were known to be impacted in PD according to previous group-wise studies. Consequently, another positive aspect of this data mining approach is that by ranking the most discriminative features within a classification framework, clinicians will be able to further investigate those top ranked features and their general clinical validity in PD diagnosis. Second, as mentioned before, non-informative and redundant features are known to downgrade classification performance. Constraining the number of the entire feature set to only

necessary input features via these methods, classification algorithms can perform the differentiation process more efficiently.

Overall, a wide range of classification techniques namely, variations of SVM such as radial basis, linear, and least square SVM, different types of decision trees, stepwise, linear, and logistic regression, MLPs, and others were used in the studies related to this research thesis. A general pattern that was observed is that linear kernel SVMs perform consistently well in PD sub-syndrome differentiation. This is particularly thought provoking, since linear SVM are better suited to handle unbalanced datasets, which happens to be the case with many of the reviewed studies here and they are less prone to over-fitting issues if a linear kernel is used compared to many other classifiers.[132] Another interesting aspect of SVMs is that multi-class variations of this classifier are well researched and available, further promoting their usage in PD-CAD systems where more than two subtypes are under investigation. In the present study, the top performing morphological classifier was a linear kernel and inherently multi-class version of SVM. Moreover, decision tree-based models such as the logistic model tree have been shown to perform remarkably well in classification tasks as they are considered state of the art.[119] LMTs, and other decision tree models have several advantages that make them invaluable in CAD-PD devices. First, due to the way decision trees are designed, they do not necessarily require feature selection as the top few nodes in a built tree essentially denote the most discriminatory features within any given classification task. This is a very important feature of decision trees as the time spent on feature selection algorithms can be directed towards optimizing the nodes and other operational parameters. Second, decision trees are less affected by outliers, missing values in training datasets, and require no data normalization.[133] Finally, unlike most classification

methods, decision trees are easier to interpret by humans. Considering that CAD-PD devices will eventually be used by clinicians, this advantage is remarkably important. While decision trees have numerous benefits, they have some downsides too. Most importantly, decision trees typically fail to recognize non-linear relationships between features. One might be tempted to implement feature selection methods such as RELIEFF in combination with decision trees to compensate for this shortcoming. In this research thesis, two out of the four top performing classification accuracies were achieved by a LMT model. This is an important finding that illustrates the potential benefits of decision trees in CAD-PD architecture. Furthermore, with the advent or rather the resurfacing of neural network such as MLPs and deep learning approaches, the operational validity of such methods in PD-CAD are future research avenues that need to be explored. MLPs have multiple advantages such as ease of use, accounting for linear and non-linear relationships, ability to handle extremely complex problems by adding multiple hidden layers and other benefits. However, the downside of MLPs include the need for large training samples and uninterpretable inner structure, which could hamper regulatory approval for a real clinical application.[134] In this work, a classification accuracy of 95.14% using an SVM based feature selection method combined with an MLP classifier was achieved. The obtained accuracy is especially encouraging considering that only three hidden layers were used in the MLP architecture. A more complex hidden layer structure combined with a large training dataset could potentially achieve even higher levels of accuracy. Furthermore, since MLPs are able to handle complex multi-class problems as well as advances in computation speed, it seems likely that neural networks and deep learning approaches are appealing methods in future CAD-PD devices.

Limitations

Three major limitations are present in this research thesis. First, the study cohort used in this work, while relatively large compared to similar studies, is still not large enough to fully expand on the generalizability of the proposed model. This limitation is further perpetuated by the lower incidences for PSP-RS compared to PD. Furthermore, an independent validation dataset, preferably acquired in different imaging centers, would be a more rigorous approach of model verification. However, this separate dataset was not available for this present study to further test the proposed model. I opted not to separate the current dataset into completely separate training and validation sub groups as the training cohort would not have been sufficiently large enough to train a generalized classifier, potentially resulting in an over-fitted model. Extra precautions such as applying the leave-one-out cross validation and the permutation testing were used to minimize the risk of over-fitting as much as possible. It is worth noting that studies employing separate validation datasets are rather scarce in this context, so that cross validation methods are used most frequently for classifier validation. Second, the ground truth classifications were determined by an expert clinician according to established consensus criteria without neuropathological proofs. Thus, there may be still a minor level of uncertainty left regarding the ground truth classification used for training and evaluation of the classifier. Third, the diseased groups differed considerably regarding the age and gender distribution, which might bias the results obtained in this research project towards higher accuracies. Ideally, we prefer to use training datasets form subjects with similar age and gender distributions in order to obtain more accurate classification models specific to age and gender. However, this is not an easy task as obtaining data with such specifications is time-consuming and expensive and is an undertaking

that was not feasible for this research project due to the retrospective data analysis. Fortunately, with the advances of medical image analysis and computer-based classification methods, it seems that such classification models will be attainable in the near future.

Future Directions

There are three broad areas of progression that will potentially build upon the concepts presented in this research thesis. The first potential direction is strongly related to the main hypothesis of this research thesis that inclusion of multi-modal MRI features would result in better classification accuracy in CAD-PD device compared to single modality approaches. Expanding on the idea of multi-source feature inclusion, we may also incorporate various other image and even non-image-based features related to the PD syndrome to create an even more comprehensive CAD-PD device. Within this context, it should also be noted that the potential of image and non-image-based biomarkers for classification of Parkinsonian syndromes is not fully exploited yet. Several previous studies have shown that other image-based biomarkers such as SWI, QSM, and others [118,135] that are not used for classification so far, provide differentiation between the diseased groups. Although parameters that differ significantly in group-wise statistics are not necessarily valuable for classification of single subjects using machine learning methods, it is still likely that these parameters may have discriminative power potentially complementary to the parameters already used for individual-level classification. Moreover, features from non-image-based biomarkers such as blood tests,[136] handwriting analysis,[137] and PD voice analysis,[138] and gait analysis[139] may also be incorporated within future CAD-PD devices to improve the comprehensiveness of the model. Moreover, additional features that can

be included in the overall feature list are clinical criteria scores. In fact, the incorporation of cognitive and neuropsychiatric symptom test scores and patient demographics may further increase the diagnostic accuracy of an automatic CAD-PD device.

The second direction is the inclusion of other PD sub-syndromes such as multiple system atrophy, corticobasal degeneration, and others for the classification model. In this thesis, only HC, PD, and PSP-RS were included, however the classification framework presented here can be easily extended if relevant sub-syndrome datasets are available. The lack of adequate numbers of PD- sub-syndrome datasets is a challenge facing many groups in the PD community. This challenge rises from several aspects, namely, that obtaining data with such specifications is time-consuming and expensive.

The third avenue, which is dependent on the success of the previous two directions, is the commercialization CAD-PD device for deployment in specialized and non-specialized movement disorder centers. Ultimately, the advancements in drug discovery for PD sub-syndromes by the time such devices are introduced in health-care institutions will surely benefit public health and lessen the burden on clinicians.

6.2 Conclusion

In the present research thesis, it was hypothesized that the inclusion of multi-modal feature sets in classification methods will improve the differentiation performance of PD, PSP-RS, and healthy controls compared to when single modality features are used. As results have indicated, the top performing classification pipeline using diffusion-based features is tied with the pipeline using multi-modal features in terms of overall accuracy, however, in the latter, the specificity is improved by 10% while sensitivity is decreased by only 3%. Irrespective of how sensitivity and specificity is defined, the multi-modal feature set will result in a higher overall sensitivity and specificity compared to the single modality approaches. The results of this work reveal a significant improvement in PD vs. PSP-RS vs. HC classification and establishes new avenues in the clinical assisted diagnosis research related to PD.

REFERENCES:

1. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. L. The prevalence of Parkinson's disease: A systematic review and meta-analysis. Mov. Disord. **29,** 1583–1590 (2014).

2. Sharma, S. et al. Biomarkers in Parkinson's disease (recent update). Neurochem. Int. **63,** 201–229 (2013).

3. Fearnley, J. M. & Lees, A. J. Ageing and Parkinson's disease: substantia nigra regional selectivity. Brain **114,** 2283–2301 (1991).

4. Gattellaro, G. et al. White matter involvement in idiopathic Parkinson disease: A diffusion tensor imaging study. Am. J. Neuroradiol. **30,** 1222–1226 (2009).

5. Tambasco, N. et al. Magnetization transfer changes of grey and white matter in Parkinson's disease. Neuroradiology **45,** 224–30 (2003).

6. Schapira, A. H. & Jenner, P. Etiology and pathogenesis of Parkinson's disease. Mov. Disord. **26,** 1049–1055 (2011).

7. Connolly, B. S. & Lang, A. E. Pharmacological treatment of Parkinson disease: A review. JAMA - J. Am. Med. Assoc. **311,** 1670–1683 (2014).

8. Yekhlef, F. et al. Routine MRI for the differential diagnosis of Parkinson's disease, MSA, PSP, and CBD. J. Neural Transm. **110,** 151–169 (2003).

9. Miller, I. N. & Cronin-Golomb, A. Gender differences in Parkinson's disease: clinical characteristics and cognition. Mov. Disord. (2010). doi:10.1002/mds.23388

10. Van Den Eeden, S. K. Incidence of Parkinson's Disease: Variation by Age, Gender, and Race/Ethnicity. Am. J. Epidemiol. **157,** 1015–1022 (2003).

11. Kaat, L. D. et al. Frontal presentation in progressive supranuclear palsy. Neurology **69,** 723–729 (2007).

12. Shen, K., Almarzouqi, S. J. & Lee, A. G. Progressive Supranuclear Palsy. Encycl. Ophthalmol. 1–3 (2015). doi:10.1007/978-3-642-35951-4_1302-1

13. Höglinger, G. U. et al. Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. Mov. Disord. **32,** 853–864 (2017).

14. Dickson, D. W. Neuropathologic differentiation of progressive supranuclear palsy and corticobasal degeneration. J. Neurol. **246,** II6-II15 (1999).

15. Lamb, R., Rohrer, J. D., Lees, A. J. & Morris, H. R. Progressive Supranuclear Palsy and Corticobasal Degeneration: Pathophysiology and Treatment Options. Curr. Treat. Options Neurol. **18,** (2016).

16. Se, D. Parkinson's Disease Society Brain Bank, London: overview and research. J Neural Transm Suppl **39,** 165–72 (1993).

17. Hachinski, V. et al. National Institute of Neurological Disorders and Stroke-Canadian Stroke Network vascular cognitive impairment harmonization standards. Stroke **37,** 2220–2241 (2006).

18. Hughes, A. J., Daniel, S. E., Kilford, L. & Lees, A. J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J. Neurol. Neurosurg. Psychiatry **55,** 181–184 (1992).

19. Kalia, L. V. & Lang, A. E. Parkinson's disease. Lancet **386,** 896–912 (2015).

20. Gröger, A. et al. Differentiation between idiopathic and atypical parkinsonian syndromes using three-dimensional magnetic resonance spectroscopic imaging. J. Neurol. Neurosurg. Psychiatry **84,** 644–9 (2013).

21. Zanigni, S. et al. Accuracy of MR markers for differentiating Progressive Supranuclear Palsy from Parkinson's disease. NeuroImage Clin. **11,** 736–742 (2016).

22. Singh, N., Pillay, V. & Choonara, Y. E. Advances in the treatment of Parkinson's disease. Prog. Neurobiol. **81,** 29–44 (2007).

23. Davie, C. A. A review of Parkinson's disease. Br. Med. Bull. **86,** 109–127 (2008).

24. Messina, D. et al. Patterns of brain atrophy in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. Parkinsonism Relat. Disord. **17,** 172–176 (2011).

25. Focke, N. K. et al. Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic Parkinson syndrome and healthy controls. Hum. Brain Mapp. **32,** 1905–1915 (2011).

26. Duchesne, S., Rolland, Y. & Verin, M. Automated computer differential classification in parkinsonian syndromes via pattern analysis on MRI. Acad. Radiol. **16,** 61–70 (2009).

27. Worker, A. et al. Cortical thickness, surface area and volume measures in Parkinson's disease, multiple system atrophy and progressive supranuclear palsy. PLoS One **9,** 1–15 (2014).

28. Price, S. et al. Voxel-based morphometry detects patterns of atrophy that help differentiate progressive supranuclear palsy and Parkinson's disease. Neuroimage **23,** 663–669 (2004).

29. Gama, R. L. et al. Morphometry MRI in the differential diagnosis of parkinsonian syndromes. Arq. Neuropsiquiatr. **68,** 333–338 (2010).

30. Quattrone, A. et al. MR imaging index for differentiation of progressive supranuclear palsy from Parkinson disease and the Parkinson variant of multiple system atrophy. Radiology **246,** 214–221 (2008).

31. Menke, R. A. L. et al. Comprehensive morphometry of subcortical grey matter structures in early-stage Parkinson's disease. Hum. Brain Mapp. **35,** 1681–1690 (2014).

32. Ulla, M. et al. Is R2* a New MRI Biomarker for the Progression of Parkinson's Disease? A Longitudinal Follow-Up. PLoS One **8,** 1–8 (2013).

33. Wood, J. C. et al. MRI R2 and R2* mapping accurately estimates hepatic iron concentration in transfusion-dependent thalassemia and sickle cell disease patients. Blood **106,** 1460–1465 (2005).

34. Alústiza Echeverría, J. M., Castiella, A. & Emparanza, J. I. Quantification of iron concentration in the liver by MRI. Insights into Imaging (2012). doi:10.1007/s13244-011-0132-1

35. Wang, Y. et al. Different iron-deposition patterns of multiple system atrophy with predominant parkinsonism and idiopathetic Parkinson diseases demonstrated by phase-corrected susceptibility-weighted imaging. Am. J. Neuroradiol. **33,** 266–273 (2012).

36. Lee, J. H. et al. Quantitative assessment of subcortical atrophy and iron content in progressive supranuclear palsy and parkinsonian variant of multiple system atrophy. J. Neurol. **260,** 2094–2101 (2013).

37. Du, G. et al. Combined diffusion tensor imaging and apparent transverse relaxation rate differentiate Parkinson disease and atypical parkinsonism. Am. J. Neuroradiol. **38,** 966–972 (2017).

38. Stejskal, E. O. & Tanner, J. E. Spin diffusion measurements: Spin echoes in the presence of a time - dependent field gradient. J. Chem. Phys. **288,** 288–292 (2014).

39. Alba-Ferrara, L. M. & de Erausquin, G. A. What does anisotropy measure? Insights from increased and decreased anisotropy in selective fiber tracts in schizophrenia. Front. Integr. Neurosci. (2013). doi:10.3389/fnint.2013.00009

40. Iacconi, C. et al. The role of mean diffusivity (MD) as a predictive index of the response to chemotherapy in locally advanced breast cancer: A preliminary study. Eur. Radiol. (2010). doi:10.1007/s00330-009-1550-z

41. Song, S.-K. et al. Dysmyelination Revealed through MRI as Increased Radial (but Unchanged Axial) Diffusion of Water. Neuroimage **17,** 1429–1436 (2002).

42. Nicoletti, G. et al. Apparent diffusion coefficient measurements of the middle cerebellar peduncle differentiate the Parkinson variant of MSA from Parkinson's disease and progressive supranuclear palsy. Brain **129,** 2679–2687 (2006).

43. Nicoletti, G. et al. Apparent diffusion coefficient of the superior cerebellar peduncle differentiates progressive supranuclear palsy from Parkinson's disease. Mov. Disord. **23,** 2370–2376 (2008).

44. Rizzo, G. et al. Diffusion-weighted brain imaging study of patients with clinical diagnosis of corticobasal degeneration, progressive supranuclear palsy and Parkinson's disease. Brain **131,** 2690–2700 (2008).

45. Agosta, F. et al. Clinical, cognitive, and behavioural correlates of white matter damage in progressive supranuclear palsy. J. Neurol. **261,** 913–924 (2014).

46. Ito, S., Makino, T., Shirai, W. & Hattori, T. Diffusion tensor analysis of corpus callosum in progressive supranuclear palsy. Neuroradiology **50,** 981–985 (2008).

47. Tsukamoto, K. et al. Significance of apparent diffusion coefficient measurement for the differential diagnosis of multiple system atrophy, progressive supranuclear palsy, and Parkinson's disease: Evaluation by 3.0-T MR imaging. Neuroradiology **54,** 947–955 (2012).

48. Seppi, K. et al. Diffusion-weighted imaging discriminates progressive supranuclear palsy from PD, but not from the parkinson variant of multiple system atrophy. Neurology **60,** 922 LP-927 (2003).

49. Prodoehl, J. et al. Diffusion tensor imaging of Parkinson's disease, atypical parkinsonism, and essential tremor. Mov. Disord. **28,** 1816–1822 (2013).

50. Karagulle Kendi, A. T., Lehericy, S., Luciana, M., Ugurbil, K. & Tuite, P. Altered diffusion in the frontal lobe in Parkinson disease. Am. J. Neuroradiol. **29,** 501–505 (2008).

51. Erbetta, A. et al. Diffusion tensor imaging shows different topographic involvement of the thalamus in progressive supranuclear palsy and corticobasal degeneration. Am. J. Neuroradiol. **30,** 1482–1487 (2009).

52. Rolheiser, T. M. et al. Diffusion tensor imaging and olfactory identification testing in early-stage Parkinson's disease. J. Neurol. **258,** 1254–1260 (2011).

53. Zhang, Y., Wu, I. W., Tosun, D., Foster, E. & Schuff, N. Progression of regional microstructural degeneration in Parkinson's disease: A multicenter diffusion tensor imaging study. PLoS One **11,** 1–16 (2016).

54. Klein, A. et al. Evaluation of 15 nonlinear deformation algorithms applied to human brain MRI registration. New York **46,** 1–62 (2009).

55. Avants, B. B., Tustison, N. & Song, G. Advanced Normalization Tools (ANTS). 1–35 (2011).

56. Brett, M., Christoff, K., Cusack, R. & Lancaster, J. Using the talairach atlas with the MNI template. Neuroimage **13,** 85 (2001).

57. Dash, M. & Liu, H. Feature selection for classification. Intell. Data Anal. **1,** 131–156 (1997).

58. Hall, M. Correlation-based Feature Selection for Machine Learning. Methodology **21i195-i20,** 1–5 (1999).

59. Karegowda, A. G., Manjunath, A. S. & Jayaram, M. A. Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection. Int. J. Inf. Technol. Knowl. Manag. **2,** 271–277 (2010).

60. Hall, M. et al. The WEKA data mining software. SIGKDD Explor. Newsl. **11,** 10 (2009).

61. Principal, P. & Analysis, C. Probabilistic Principal Component Analysis and the EM algorithm. October (2007).

62. Kononenko, I., Šimec, E. & Robnik-Šikonja, M. Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl. Intell. **7,** 39–55 (1997).

63. Zhou, X. & Wang, J. Feature Selection for Image Classification Based on a New Ranking Criterion. J. Comput. Commun. **3,** 74–79 (2015).

64. Guyon, I. Gene selection for cancer classification using support vector machines. Mach. Learn. **46,** 389–422 (2002).

65. Quinlan, J. R. Induction of Decision Trees. Mach. Learn. **1,** 81–106 (1986).

66. Rokach, L. & Maimon, O. in Data Mining and Knowledge Discovery Handbook 165–192 (2010). doi:10.1007/0-387-25465-X_9

67. Breiman, L. (University of C. Random forest. Mach. Learn. **45,** 1–35 (1999).

68. Landwehr, N., Hall, M. & Frank, E. Logistic model trees. Mach. Learn. **59,** 161–205 (2005).

69. Peterson, L. E. K-nearest neighbor. Scholarpedia **4,** 1883 (2009).

70. Zhang, H. The Optimality of Naive Bayes. Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2004 **1,** 1–6 (2004).

71. Cortes, C. & Vapnik, V. Support vector machine. Mach. Learn. **20,** 273–297 (1995).

72. Riedmiller, M. Advanced supervised learning in multi-layer perceptrons - From backpropagation to adaptive learning algorithms. Comput. Stand. Interfaces **16,** 265–278 (1994).

73. Scherfler, C. et al. Diagnostic potential of automated subcortical volume segmentation in atypical parkinsonism. Neurology **86,** 1242–1249 (2016).

74. Fischl, B. FreeSurfer. NeuroImage **62,** 774–781 (2012).

75. Sarica, A. et al. Application of different classification techniques on brain morphological data. Comput. Med. Syst. (CBMS), 2013 IEEE 26th Int. Symp. on. IEEE 425–428 (2013). doi:10.1109/CBMS.2013.6627832

76. Salvatore, C. et al. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy. J. Neurosci. Methods **222,** 230–237 (2014).

77. Rana, B. et al. Graph-theory-based spectral feature selection for computer aided diagnosis of Parkinson's disease using T1-weighted MRI. Int. J. Imaging Syst. Technol. **25,** 245–255 (2015).

78. Friston, K. & Ashburner, J. Statistical parametric mapping. Funct. neuroimaging Tech. Found. 1–74 (1994). doi:10.4249/scholarpedia.6232

79. Singh, G. & Samavedham, L. Unsupervised learning-based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of Parkinson disease. J. Neurosci. Methods **256,** 30–40 (2015).

80. Kohonen, T. The self-organizing map. Neurocomputing **21,** 1–6 (1998).

81. Peng, B. et al. A multilevel-ROI-features-based machine learning method for detection of morphometric biomarkers in Parkinson's disease. Neurosci. Lett. **651,** 88–94 (2017).

82. Mueller, K. et al. Disease-specific regions outperform whole-brain approaches in identifying progressive supranuclear palsy: A multicentric MRI study. Front. Neurosci. **11,** 1–11 (2017).

83. Lin, W. et al. Parkinson's disease: diagnostic utility of volumetric imaging. Neuroradiology **59,** 367–377 (2017).

84. Hammers, A. et al. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. **19,** 224–247 (2003).

85. Boelmans, K. et al. Brain iron deposition fingerprints in Parkinson's disease and progressive supranuclear palsy. Mov. Disord. **27,** 421–427 (2012).

86. Eckert, T. et al. Differentiation of idiopathic Parkinson's disease, multiple system atrophy, progressive supranuclear palsy, and healthy controls using magnetization transfer imaging. Neuroimage **21,** 229–235 (2004).

87. Haller, S. et al. Differentiation between Parkinson disease and other forms of Parkinsonism using support vector machine analysis of susceptibility-weighted imaging (SWI): Initial results. Eur. Radiol. **23,** 12–19 (2013).

88. Hess, C. W., Ofori, E., Akbar, U., Okun, M. S. & Vaillancourt, D. E. The evolving role of diffusion magnetic resonance imaging in movement disorders. Curr. Neurol. Neurosci. Rep. **13,** (2013).

89. Scherfler, C. et al. Voxel-wise analysis of diffusion weighted imaging reveals disruption of the olfactory tract in Parkinson's disease. Brain **129,** 538–542 (2006).

90. Salamanca, L., Vlassis, N., Diederich, N., Bernard, F. & Skupin, A. Improved Parkinson's Disease Classification from Diffusion MRI Data by Fisher Vector Descriptors. nternational Conf. Med. Image Comput. Comput. Interv. 119–126 (2015). doi:10.1007/978-3-319-24553-9

91. Perronnin, F., Liu, Y., Sánchez, J. & Poirier, H. Large-scale image retrieval with compressed fisher vectors. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 3384–3391 (2010). doi:10.1109/CVPR.2010.5540009

92. Haller, S. et al. Individual detection of patients with Parkinson disease using support vector machine analysis of diffusion tensor imaging data: initial results. AJNR. Am. J. Neuroradiol. **33,** 2123–8 (2012).

93. Banerjee, M., Okun, M. S., Vaillancourt, D. E. & Vemuri, B. C. A method for automated classification of Parkinson's disease diagnosis using an ensemble average propagator template brain map estimated from diffusion MRI. PLoS One **11,** 1–11 (2016).

94. Moriarty, T. F. Two sets of inequalities among the principal invariants of the Cauchy-Green deformation tensors. J. Elast. **1,** 87–90 (1971).

95. Zhang, M. & Fletcher, P. T. Probabilistic Principal Geodesic Analysis. Adv. Neural Inf. Process. Syst. 26 1178–1186 (2013).

96. Péran, P. et al. Magnetic resonance imaging markers of Parkinson's disease nigrostriatal signature. Brain **133,** 3423–3433 (2010).

97. Morisi, R., Cha, M., Arafa, M. & Zagrouba, E. Binary and Multi-class Parkinsonian Disorders Classification Using Support Vector Machines. Iber. Conf. Pattern Recognit. Image Anal. 379–386 (2015). doi:10.1007/978-3-319-19390-8

98. Oguz, I. et al. DTIPrep: quality control of diffusion-weighted images. Front. Neuroinform. **8,** (2014).

99. Planetta, P. J. et al. Free-water imaging in Parkinson's disease and atypical parkinsonism. Brain **139,** 495–508 (2016).

100. Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. J. Am. Geriatr. Soc. **53,** 695–699 (2005).

101. Cherubini, A. et al. Magnetic resonance support vector machine discriminates between Parkinson disease and progressive supranuclear palsy. Mov. Disord. **29,** 266–9 (2014).

102. Nemmi, F., Sabatini, U., Rascol, O. & Péran, P. Parkinson's disease and local atrophy in subcortical nuclei: Insight from shape analysis. Neurobiol. Aging **36,** 424–433 (2015).

103. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. Neuroimage (2012). doi:10.1016/j.neuroimage.2011.09.015

104. Ota, M. et al. Differential diagnosis tool for parkinsonian syndrome using multiple structural brain measures. Comput. Math. Methods Med. **2013,** (2013).

105. Long, D. et al. Automatic Classification of Early Parkinson's Disease with Multi-Modal MR Imaging. PLoS One **7,** 1–9 (2012).

106. Tzourio-Mazoyer, N. et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage (2002). doi:10.1006/nimg.2001.0978

107. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. & Baker, C. I. Circular analysis in systems neuroscience: The dangers of double dipping. Nat. Neurosci. **12,** 535–540 (2009).

108. Tolosa, E., Wenning, G. & Poewe, W. The diagnosis of Parkinson's disease. Lancet Neurol. **5,** 75–86 (2006).

109. Litvan, I. et al. Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome): Report of the NINDS-SPSP International Workshop. Neurology **47,** 1–9 (1996).

110. Forkert, N. D., Cheng, B., Kemmling, A., Thomalla, G. & Fiehler, J. ANTONIA perfusion and stroke: A software tool for the multi-purpose analysis of MR perfusion-weighted

datasets and quantitative ischemic stroke assessment. Methods Inf. Med. **53,** 469–481 (2014).

111. Mazziotta, J. et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos. Trans. R. Soc. B Biol. Sci. **356,** 1293–1322 (2001).

112. Ourselin, S., Roche, A., Subsol, G., Pennec, X. & Ayache, N. Reconstructing a 3D structure from serial histological sections. Image Vis. Comput. **19,** 25–31 (2001).

113. Modat, M. et al. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. **98,** 278–284 (2010).

114. B. B. Avants, C. L. Epstein, M. Grossman, J. C. G. Symmetric Diffeomorphic Image Registration with Cross- Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain. Med Image Anal **12,** 26–41 (2008).

115. Lan, J. & Jiang, D. H. Excessive iron accumulation in the brain: A possible potential risk of neurodegeneration in Parkinson's disease. J. Neural Transm. **104,** 649–660 (1997).

116. Cochrane, C. J. & Ebmeier, K. P. Diffusion tensor imaging in parkinsonian syndromes. Neurology **80,** 857 LP-864 (2013).

117. Dabrowska, M. et al. The utility of susceptibility-weighted imaging for differentiating Parkinsonism-predominant multiple system atrophy from Parkinson's disease: Correlation with 18F-flurodeoxyglucose positron-emission tomography. Brain Res. **9,** 60–3 (2015).

118. Murakami, Y. et al. Usefulness of quantitative susceptibility mapping for the diagnosis of Parkinson disease. Am. J. Neuroradiol. **36,** 1102–1108 (2015).

119. Singh, G. & Samavedham, L. Algorithm for image-based biomarker detection for differential diagnosis of Parkinson's disease. IFAC-PapersOnLine **48,** 918–923 (2015).

120. Han, H. & Jiang, X. Overcome support vector machine diagnosis overfitting. Cancer Inform. **13,** 145–58 (2014).

121. Rana, B. et al. Regions-of-interest based automated diagnosis of Parkinson's disease using T1-weighted MRI. Expert Syst. Appl. **42,** 4506–4516 (2015).

122. Schwarz, S. T. et al. T1-Weighted MRI shows stage-dependent substantia nigra signal loss in Parkinson's disease. Mov. Disord. (2011). doi:10.1002/mds.23722

123. Vymazal, J. et al. T1 and T2 in the brain of healthy subjects, patients with Parkinson disease, and patients with multiple system atrophy: relation to iron content. Radiology (1999). doi:10.1148/radiology.211.2.r99ma53489

124. Whitwell, J. L. et al. Disrupted thalamocortical connectivity in PSP: A resting-state fMRI, DTI, and VBM study. Park. Relat. Disord. (2011). doi:10.1016/j.parkreldis.2011.05.013

125. Boxer, A. L. et al. Patterns of brain atrophy that differentiate corticobasal degeneration syndrome from progressive supranuclear palsy. Archives of Neurology (2006). doi:10.1001/archneur.63.1.81

126. Tessitore, A. et al. Regional gray matter atrophy in patients with Parkinson disease and freezing of gait. AJNR. Am. J. Neuroradiol. **33,** 1804–9 (2012).

127. Messina, D. et al. Patterns of brain atrophy in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. Parkinsonism Relat. Disord. **17,** 172–176 (2011).

128. Nandakumar S. Narayanan, Robert L. Rodnitzky, E. U. NIH Public Access. Rev Neurosci **24,** 267–278 (2013).

129. Aerts, M. B., Abdo, W. F. & Bloem, B. R. The 'bicycle sign' for atypical parkinsonism. The Lancet (2011). doi:10.1016/S0140-6736(11)60018-4

130. Gelb, D., Oliver, E. & Gilman, S. Diagnostic criteria for Parkinson disease. J Neurol Neurosurg Psychiatry **79,** 368–376 (2008).

131. Mudali, D., Teune, L. K., Renken, R. J., Leenders, K. L. & Roerdink, J. B. T. M. Classification of Parkinsonian Syndromes from FDG-PET Brain Data Using Decision Trees with SSM / PCA Features. Comput. Math. Methods Med. **2015,** (2015).

132. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J. Mach. Learn. Res. (2010).

133. Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. Informatica (2007). doi:10.1115/1.1559160

134. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. **49,** 1225–1231 (1996).

135. Gupta, D., Saini, J., Kesavadas, C., Sarma, P. S. & Kishore, A. Utility of susceptibility-weighted MRI in differentiating Parkinson's disease and atypical parkinsonism. Neuroradiology **52,** 1087–1094 (2010).

136. DeMarshall, C. A. et al. Potential utility of autoantibodies as blood-based biomarkers for early detection and diagnosis of Parkinson's disease. Immunol. Lett. **168,** 80–88 (2015).

137. Drotar, P. et al. Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease. Comput. Methods Programs Biomed. **117,** 405–411 (2014).

138. Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J. & Ramig, L. O. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans. Biomed. Eng. **56,** 1015–1022 (2009).

139. Zeng, W. et al. Parkinson's disease classification using gait analysis via deterministic learning. Neurosci. Lett. **633,** 268–278 (2016).