Prediction of neurodegenerative diseases from functional brain imaging data

Mudali, Deborah

PREDICTION OF NEURODEGENERATIVE DISEASES FROM FUNCTIONAL BRAIN IMAGING DATA

DEBORAH MUDALI

This research was supported by the Netherlands Fellowship Programmes (NFP) of Nuffic under grant number CF6695/2010.

Cover: Three orthogonal slices of the first principal component volume of FDG-PET brain scans of Parkinson's disease subjects compared to healthy controls, overlaid on an anatomical brain template. Also shown is a decision tree diagram of the classification output of the subjects.

Mudali, Deborah

Prediction of Neurodegenerative Diseases from Functional Brain Imaging Data
Deborah Mudali
Thesis Rijksuniversiteit Groningen

ISBN 978-90-367-8694-2 (printed version)
ISBN 978-90-367-8693-5 (electronic version)

university of groningen

Prediction of Neurodegenerative Diseases from Functional Brain Imaging Data

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 14 March 2016 at 12.45 hours

by

Deborah Mudali

born on October 2nd, 1982
in Iganga, Uganda

Supervisors
Prof. J. B. T. M. Roerdink
Prof. M. Biehl

Assessment committee
Prof. B. M. ter Haar Romeny
Prof. N. M. Maurits
Prof. A. C. Telea

I dedicate this work to my dear late mother Robinah Baseke Batwaula my inspiration and my father Moses Batwaula.

CONTENTS

1 INTRODUCTION

The diagnosis of neuro-degenerative diseases characterised by slow progression is difficult, especially at an early stage. These diseases have continued to affect the elderly [Berg, 2008], especially in developed countries where life expectancy is high. Some of these disorders include Parkinson's disease (PD), progressive supranuclear palsy (PSP), multisystem atrophy (MSA), Alzheimer's disease (AD), frontotemporal dementia (FTD), and dementia with Lewy bodies (DLB), to mention a few. Parkinson's disease (PD) is a progressive disorder which causes slow motion and rigidity in the body. PD is characterized by neuronal loss in the substantia nigra and other brain regions, also associated with the formation of intracellular protein inclusions known as Lewy bodies [Shulman and De Jager, 2009]. On the other hand, Alzheimer's disease (AD) is associated with progressive memory loss, as well as judgment and decision making impairments, according to statistics collected by Guttmacher et al. [2003].

There is increasing interest to use neuroimaging techniques in the hope to discover biomarkers, that is, abnormal patterns of morphology, energy consumption or network activity of the brain which are characteristic for such diseases. Many *in vivo* brain imaging techniques are nowadays available for this purpose; see Table 1.1. Positron Emission Tomography (PET) has been applied in many medical studies due to its capability to show metabolism of the brain. PET scans reveal the amount of metabolic activity in various parts of the brain. PET is used for clinical diagnosis of brain diseases. For example, in patients with idiopathic Parkinson's disease (IPD), a pattern of increased metabolism in specific brain areas was found on the basis of Positron Emission Tomography (PET) imaging [Fodero-Tavoletti et al., 2009]. This method is based on visual comparison of patient data with disease-specific templates that contain regions of interest (ROIs), and is therefore quite subjective. Other studies to detect and differentiate parkinsonian syndromes include Eckert et al. [2005, 2008]; Hellwig et al. [2012]; Wu et al. [2013]; Garraux et al. [2013]. On the other

hand, PET has been used in previous studies to test for depression and dementia in patients with neurodegenerative diseases. For example, as reported by Chow et al. [2009], PET was used to measure the availability of cortical serotonin receptors in older subjects. PET can also be used to perform efficient diagnosis [Yun et al., 2001]. Newly developed PET techniques use multivariate statistical analysis to identify disease-related network patterns of brain metabolism [Frackowiak et al., 2003].

Although this thesis focuses on PET for image acquisition, we briefly mention some other methods, also because some of these may be combined with PET in future work. A large collection of brain imaging techniques is based upon Magnetic Resonance Imaging (MRI) [Golder, 2002]. In contrast to PET, which requires administering radioactive tracers to the subject being scanned, MRI is a fully non-invasive method with no known health implications. In fact, distinct spatial patterns of cortical atrophy have been found from structural MRI images by a technique called voxel-based morphometry (VBM) [Chételat et al., 2008; Berg, 2008]. This method considers all voxels in a brain volume, and gives quantitative estimates of the grey and white matter volumes without assuming any *a priori* regions of interest. Another MRI-based technique, called Diffusion Tensor Imaging (DTI), is able to measure the amount of anisotropy of water diffusion, from which the orientation of nerve fiber bundles in brain white matter can be inferred [Basser et al., 1994]. Also, in the study by Ito et al. [2006] measures like apparent diffusion coefficient (ADC) and fractional anisotropy (FA) have been used to evaluate the degree of tissue degeneration in diseases like Parkinson's and multiple system atrophy. In addition, DTI-tractography can be used to visualise nerve fiber tracts and study abnormal connectivity patterns between brain regions. Furthermore, MRI-based techniques such as Arterial Spin Labelling (ASL) and Susceptibility-Weighted Imaging (SWI) allow for quantitative assessment of tissue perfusion and levels of venous blood, hemorrhage, and iron storage in the brain, respectively. Additionally, the MRI-technique known as functional magnetic resonance imaging (fMRI) determines and visualises changes in brain activity that are elicited by asking test persons to carry out specific cognitive or sensorimotor tasks. A recent addition is "resting-state" fMRI, where the subject in the MRI-scanner is imaged without external stimulus; the data are processed to derive brain connectivity patterns which are assumed to represent a default-mode network [Salvador et al.,

2005] and other networks mentioned by Van Den Heuvel and Pol [2010]. This network includes regions that are known to be impaired in certain types of neuro-degenerative diseases [Greicius, 2008]. For completeness, we also have listed some imaging techniques in Table 1.1 which can detect neuronal activity, like electroencephalography (EEG) and magnetoencephalography (MEG).

Table 1.1: *In vivo* brain imaging techniques. See Figure 1.1 for a PET brain image example.

Technique	*Represents*	*Physical effect*
CT	anatomy	X-ray attenuation
PET	metabolism	radioactive decay
MRI	anatomy	magnetic resonance
fMRI	metabolism	blood deoxygenation
DTI	nerve fibers	anisotropic diffusion
MEG	neuronal activity	magnetic signals
EEG	neuronal activity	electric signals

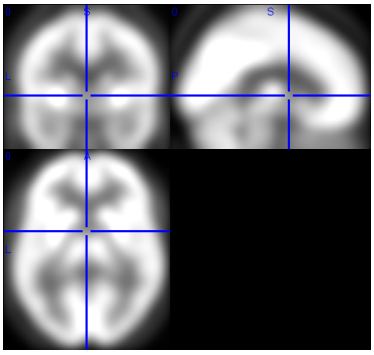


Figure 1.1: In vivo brain imaging techniques: example of a PET brain image.

Although some success has been obtained by the techniques mentioned above, a major problem is that each of the separate brain imaging modalities only produces a clearly observable disease-related brain pattern when the disease has reached an advanced stage. Also, abnormal patterns may not be specific for a single brain disease. Functional imaging methods (PET, fMRI) may give only a partial picture due to compensatory mechanisms in patients with neuro-degenerative diseases. It

is increasingly recognised that combining information derived from different image modalities is essential for improvement of the sensitivity and specificity of proposed biomarkers for neuro-degenerative diseases [Davatzikos et al., 2008]. Another shortcoming of current efforts is that only very few studies in the literature report abnormal patterns for a given disease, so that reproducibility of these findings is not yet firmly established. Also, the typical size (10-20) of patient groups is too small to differentiate between subtypes of a given disease (such as akinetic-rigid versus tremor-dominant PD). It is clear that progress in the early diagnosis of neuro-degenerative diseases can only be made if both imaging and diagnostic data of large numbers of patients in several phases of disease progression are accumulated. Also, for the many MRI-based techniques further improvements may be possible by optimising scanner sequences for each modality, which therefore have to be carefully recorded for each scan session. This will require substantial efforts in database formation during longitudinal studies spanning several decades.

1.1 OBJECTIVE

The objective of this thesis is, first, to derive features from medical imaging data in the hope to discover more sensitive and specific biomarkers for the prediction of neuro-degenerative diseases. Second, to develop supervised classification methods for associating brain patterns (and features) extracted from multi-modal brain imaging data to various types and stages of neuro-degenerative diseases.

1.1.1 *Specific Objectives*

- To collect medical data and use it to identify structural and functional brain patterns that display significant differences between healthy subjects and patients with neuro-degenerative diseases.
- To develop classification methods based on the brain features extracted from the multi-modal brain images.
- To test the performance of the developed methods.

1.2 TECHNIQUES AND TOOLS

1.2.1 *Imaging Data Acquisition by Positron Emission Tomography*

In this thesis PET was used for image acquisition. PET is an imaging technique that uses radioactive material to diagnose diseases. The radiotracer is injected into a patient, where it accumulates in the body to the specific part of interest. The tracer emits positrons which are annihilated by electrons under emission of gamma rays. The gamma rays are detected by a device called a gamma camera. The device works with the computer to measure the position-dependent amount of radiotracer absorbed by the body, thus producing a brain image.

LIMITATIONS OF PET

- The scan session takes quite a long time (approximately 30 to 40 minutes [Carne et al., 2007]) in addition to time reserved for the radiotracer (45 to 60 minutes [Townsend, 2004]) to accumulate in the body part of interest. The duration of the whole scan process also depends on the tissue under observation.
- The PET scan may show false results in case of chemical imbalances in the body.

1.2.2 *Analysis Tools*

A number of analysis tools are briefly mentioned here for completeness because they are used to pre-process the PET brain images [Teune et al., 2013], but are not further discussed in this thesis.

1.2.2.1 *Statistical Parametric Mapping (SPM)*

The SPM package [Friston et al., 2007] can be used to process images for feature extraction. Processing of images can include segmentation, co-registration and normalisation, all which are found in the SPM package.

1.2.2.2 *Voxel Based Morphometry (VBM)*

VBM is a very useful method to analyse images on the voxel level Ashburner and Friston [2000]. For example, in the study by Chételat et al. [2008], VBM was used to determine the difference in levels or stages of progression of NDs like Alzheimer's. VBM has also been used to differentiate several NDs based on pathology. Burton et al. [2004] used VBM to study the pattern of cerebral atrophy in PD in comparison to healthy controls, AD and DLB; the result was that atrophy is greater in some brain areas for AD than for PD. Generally VBM has been used for differentiation of several NDs and their stages.

1.2.3 *Classification Tools/Pattern classification*

Supervised classification methods for associating brain patterns to various forms of neuro-degenerative disease will be applied to PET data in this thesis. By using a data set of subjects whose diagnosis has been determined by medical experts, a classifier is trained using the set of features in conjunction with subject labels. After training, the classifier can be used to classify new subjects for which their diagnosis is unknown. Extracted features of the new subject are compared with the training set features indirectly. That is to say, using the rules established by the classifier during the training, the classifier computes the best matching disease type(s). This procedure is visualised in Figure 1.2. The training phase of the classification system is shown on the left. The classification or query phase, shown on the right, employs the same procedures, the features are extracted from the test case(s) and used for classification. That is, the label(s) of the test/query image(s) is determined.

An important requirement for this method to work is that the number of training samples per disease (sub)type is sufficiently high. As more training data become available in the course of time the classifiers are expected to increase in classification performance.

There are many different classification methods in neural-network and statistical-decision theory. Within learning and reasoning approaches, decision trees (DT) are among the most popular ones [Breiman et al., 1984; Quinlan, 1993]. They are also intuitive for non-experts, since the DT procedure corresponds to the way humans perform classification based on many fea-

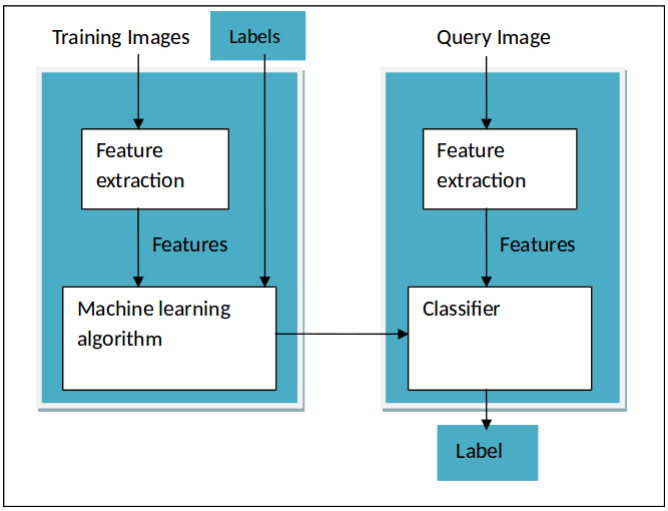


Figure 1.2: Classification System: On the left is the training phase and on the right is the testing phase.

tures (like in taxonomy). Therefore we start our investigations with this classification method. After DT, we concentrate on linear classifiers (with linear decision boundary) because more complex systems (multilayered neural networks, SVM with non-linear kernel etc.) might be at risk to over-fit the relatively small datasets. Distance based methods such as Generalized Matrix Learning Vector Quantization (GMLVQ) will be applied and compared with the decision tree algorithms. Feature reduction/definition is an integral part of these methods [Schneider et al., 2007]. Alternative classifiers such as the Support Vector Machine (SVM) [Hammer et al., 2004] will be applied as well.

1.2.3.1 *Decision trees*

A decision tree represents a recursive partition of the instance space [Rokach and Maimon, 2010]. It consists of at least a root node which can be connected by successive edges to child nodes. These child nodes, also known as internal nodes, are in turn connected to child nodes, until the leaf nodes are reached which do not have out-going edges. A new data example is classified by going through a path from the root node to the leaf node, while testing each feature of that example represented at each of the internal nodes. Based on the outcome of each test, a sequence of edges is followed until a leaf node is reached.

Since each leaf carries a class label, the new data example is assigned the class of the leaf it reaches. There are algorithms that can be used to construct decision trees which include C4.5 [Quinlan, 1993], CART Breiman et al. [1984] and others. In particular, the C4.5 decision tree inducer uses an information theoretic criterion to build decision trees. A dataset is split into subsets at each node by choosing the attribute/feature that maximizes the information gain. The details about information gain are found in chapter 3. The optimal decision tree is the one which minimizes the generalization error. Increased robustness is provided by applying "bagging" [Breiman, 1996]. For the problem considered here, i.e., brain images which would require human interpretation, a decision tree-based approach is very suitable, because it resembles the way that human experts perform classification.

1.2.3.2 *Generalized Matrix Learning Vector Quantization*

GMLVQ estimates the relevance of features in their ability to classify data. Then the classifier uses the weighted features (according to their relevance) and class prototypes to separate groups of data. This is possible with the full matrix Λ which accounts for pairwise correlations of the feature dimensions. A distance metric is used that has the form $d^{\Lambda}(\boldsymbol{w_k}, \boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{w_k})^T \Lambda (\boldsymbol{x} - \boldsymbol{w_k})$, where Λ is a positive semi-definite $N \times N$ matrix which is used to quantify the dissimilarity of an input vector $\boldsymbol{x}$ and the prototypes $\boldsymbol{w_k}$ [Schneider et al., 2009].

1.2.3.3 *Support Vector Machine*

The Support vector machine determines the optimal hyperplane with the largest distance or margin between support vectors (border-line training data examples) separating the instances in the feature space. A new data example is classified as belonging to either of the classes separated by the hyperplane. For example, given training data with input vectors $x_1, x_2, ..., x_n \in R^d$ and labels $y_1, y_2, ..., y_n \in \{-1, +1\}$ [Oz and Kaya, 2013], as shown in Figure 1.3, we need to find an optimal hyperplane $w \cdot x + b = 0$ (i.e., vector w and the scalar b) which separates the negative data examples from the positive data examples. There could exist a number of such hyperplanes, but SVMs find the hyperplane that maximizes the gap between the support vectors. This gap (as seen in Figure 1.3) is the

distance between parallel hyperplanes $w \cdot x + b = -1$ and $w \cdot x + b = +1$, i.e., $\frac{2}{||w||}$. In order to maximize the gap we need to minimize $\frac{1}{2}||w||^2$ under the following constraints:

$$w \cdot x_i + b \leq -1 \text{ if } y_i = -1$$
$$w \cdot x_i + b \geq +1 \text{ if } y_i = +1$$

Equivalently:

$$y_i(w \cdot x_i + b) \geq 1, i = 1, ..., n.$$

Generally, we want to maximize $\frac{2}{||w||}$ subject to $y_i(w \cdot x_i + b) \geq 1$ for $i = 1, ..., n$. Then given a new data example x, the decision function is signum$(f(x))$, where $f(x) = w \cdot x + b, w \in R^d$ and $b \in R$.

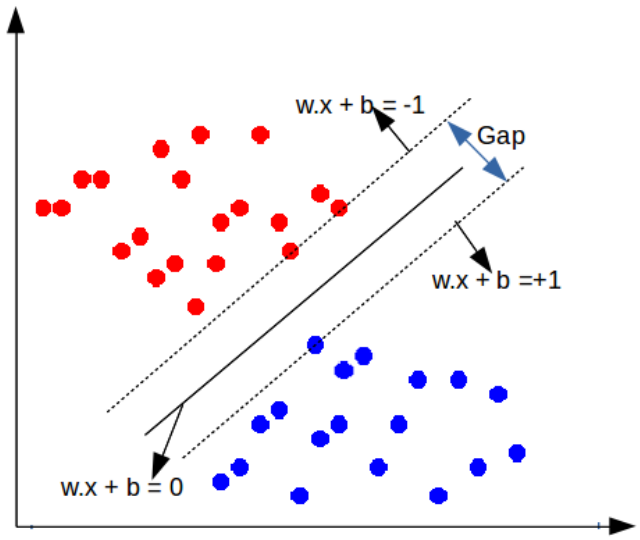


Figure 1.3: Linear SVM: The optimal hyperplane separating two classes, i.e., red dots for the negative class and blue dots for the positive class.

However, if the classes are not linearly separable, the large margin concept has to be extended in order to tolerate misclassifications Cristianini and Shawe-Taylor [2000].

1.2.4 *Visualization Tools*

In addition to the pattern recognition algorithms, we will also apply visualisation techniques such as scatter plot matrices [Zheng et al., 2014], decision tree diagrams [Stiglic et al., 2012], or multiple views [Wang Baldonado et al., 2000] to explore the labeled data sets in feature space. Visualisation methods will serve two important goals. First, they will give an insight in the distribution of the data points in the feature space, hence portraying an idea of how data can be separated into distinct classes. Second, visualisation allows an intuitive way to present the results to the medical experts, thereby facilitating communication.

1.3 SSM/PCA METHOD FOR FEATURE EXTRACTION

The scaled subprofile model with principal component analysis (SSM/PCA) [Moeller et al., 1987; Moeller and Strother, 1991] is used in this thesis to extract patterns from PET data with the removal of the group mean and the voxel mean, thereby removing the overall major subject and group global effects before applying PCA to the data. This process makes evident the main patterns of metabolic brain activity in the data. It is from these patterns that the features to be used in the classification process are determined.

In this thesis, the extracted features depend entirely on the whole input dataset, since they are produced by PCA. This makes the leave-one-out method used for performance evaluation more complicated than usual. In other words, since the features are dependent on the input data, for each leave-one-out run, a test subject is always removed from the training set before applying the SSM/PCA method and later projected onto the extracted patterns (from the training set) to obtain its scores.

1.4 THESIS CONTRIBUTION AND CONTENT

Machine learning methods are employed in the classification of neurodegenerative diseases. In chapter 2, which is based on [Mudali et al., 2015], we look at the classification of parkinsonian syndromes since they are not easily separable. We used

the C4.5 decision tree inducer [Quinlan, 1993] to train classifiers on the subject scores as features extracted from FDG-PET data.

Having applied a different method called stepwise regression (SR) [Teune et al., 2013] to the same parkinsonian syndromes data, we studied the difference between this method and the decision tree (DT) method. This is the topic of chapter 3, which is based on [Mudali et al., 2016c].

Other classification methods were introduced in chapter 4 in the hope to improve classification accuracy. This chapter is based on [Mudali et al., 2016b]. The GMLVQ and SVM classifiers were trained using features extracted from the FDG-PET data, similar to the method used in chapter 2 and chapter 3. The same SSM/PCA method was applied to the FDG-PET data to extract the features, specifically subject scores. These subject scores were input to the GMLVQ and SVM classifiers to determine the correct subject(s) label(s). Using leave-one-out cross validation, the classifier performances were evaluated.

In chapter 5, which is based on [Mudali et al., 2016a], more PD data consisting of later disease stage brain images was acquired and combined with the early stage data. The three classification methods i.e., decision trees, GMLVQ and SVM, were applied to combinations of the early and late-stage datasets. Additionally, we interchanged the later and earlier disease stage datasets for training and testing the classifiers.

Lastly, chapter 6 contains a summary of the thesis and possibilities for future work.

2

CLASSIFICATION OF PARKINSONIAN SYNDROMES FROM FDG-PET BRAIN DATA USING DECISION TREES WITH SSM/PCA FEATURES

ABSTRACT: *Medical imaging techniques like fluorodeoxyglucose positron emission tomography (FDG-PET) have been used to aid in the differential diagnosis of neurodegenerative brain diseases. Visual Interpretation of FDG-PET scans and clinical symptoms of patients with neurodegenerative brain diseases can be difficult, especially at an early disease stage. In this study, the objective is to classify FDG-PET brain scans of subjects with parkinsonian syndromes (Parkinson's disease, Multiple System Atrophy, and Progressive Supranuclear Palsy), compared to healthy controls. The scaled subprofile model/principal component analysis (SSM/PCA) method was applied to FDG PET brain image data to obtain covariance patterns and corresponding subject scores. The latter were used as features for supervised classification by the C4.5 decision tree method. Leave-one-out cross validation was applied to determine classifier performance. We carried out a comparison with other types of classifiers. The performance of the decision tree method is in some cases (somewhat) lower than that of other classifiers like nearest neighbors or support vector machines. However, the big advantage of decision tree classification is that the results are easy to understand by humans. A visual representation of decision trees strongly supports the interpretation process, which is very important in the context of medical diagnosis. Further improvements are suggested based on enlarging the number of the training data, enhancing the decision tree method by bagging, and adding additional features based on (f)MRI data.*

KEYWORDS: *Parkinsonian syndromes, FDG-PET data, scaled subprofile model, principal component analysis, decision tree classification, visual analysis.*

2.1 INTRODUCTION

Neurodegenerative brain diseases like Parkinson's disease (PD), multiple system atrophy (MSA), or progressive supranuclear palsy (PSP), are difficult to diagnose at early disease stages [Litvan et al., 2003]. It is important to develop neuroimaging techniques that can differentiate among the various forms of parkinsonian syndromes and stages in progression. Early disease detection is aided by brain imaging techniques like [18F]-fluorodeoxyglucose (FDG) positron emission tomography

(PET) and magnetic resonance imaging (MRI), to obtain image data and derive significant patterns of changed brain activity. Several techniques have been developed to identify disease-related network patterns of cerebral glucose metabolism.

Covariance techniques like principal component analysis (PCA) can be used to extract significant patterns from brain image data. PCA is known for its capability to identify patterns in high-dimensional data like brain image data. A possible approach to biomarker identification is the scaled subprofile model/principal component analysis (SSM/PCA) method [Moeller et al., 1987; Moeller and Strother, 1991]. SSM/PCA is a feature extraction method which enhances identification of significant patterns in multivariate imaging data. This method has been extensively applied to positron emission tomography data to identify brain patterns which display significant differences between healthy controls and parkinsonian conditions. The SSM/PCA method helps to reduce data dimensions and to reveal the brain patterns characteristic for a certain parkinsonian syndrome. Resting state metabolic networks obtained from FDG-PET scans were used to identify disease related metabolic brain patterns of PD, MSA and PSP [Ma et al., 2007; Eckert et al., 2008; Eidelberg, 2009; Teune et al., 2013]. In a previous study by Tang et al. [2010], it was demonstrated that by using an image-based classification routine, it was possible to distinguish with high specificity between PD and MSA/PSP, and in a second step between MSA and PSP as compared to controls.

In a recent study of Hellwig et al. [2012], the diagnostic accuracy of FDG-PET in discriminating parkinsonian patients was investigated. FDG-PET scans were analyzed by visual assessment including individual voxel based statistical maps (a 3D stereotactic surface projection technique; 3D-SSP). These studies compared only two classes at a time or on two levels (healthy and patient group, or two patient groups). This puts forward a research challenge to improve the SSM/PCA method, to be able to distinguish different neurodegenerative brain diseases from each other in one analysis.

For this reason we consider machine learning approaches like decision-tree methods to be able to compare more than two patient groups at the same time and possibly detect subtypes within patient groups. The C4.5 decision tree classification algorithm by Quinlan [1993] is used to classify parkinsonian conditions from FDG-PET imaging data. This algorithm uses a

feature selection criterion known as information gain to induce decision trees from training data. The subject scores derived from the SSM/PCA method are used as input features for the C4.5 algorithm. After the training phase, the decision trees can then be used as predictors for unseen cases with unknown disease type. Decision trees are known to be intuitive and easily understandable by humans [Cintra et al., 2012]. In other words, they can be easily visualized and interpreted by the clinicians.

In this chapter, we combine the SSM/PCA method in a novel way with the C4.5 decision tree classification algorithm which classifies parkinsonian disorders according to their respective disease types. We also compare the decision tree method with a number of other classifiers with respect to different criteria, such as performance and interpretability by humans.

2.2 MATERIALS AND METHODS

The extraction of patterns and classification involves four main steps: data acquisition, feature extraction, feature selection, and classification, see Figure 2.4.

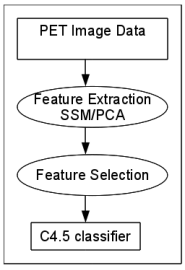


Figure 2.4: Classification steps.

2.2.1 *Data Acquisition*

FDG-PET scans from a previous study [Teune et al., 2010] describing 18 healthy controls (HC), 20 PD, 21 MSA, and 17 PSP patients were used for the present analysis. At the time of referral for imaging, the clinical diagnosis of most patients was uncertain. The final clinical diagnoses according to established clinical research criteria [Gilman et al., 2008; Litvan et al., 1996, 2003] were made after a follow-up time after scanning of 4 ± 3

years (y) in PD, 2±1y in MSA, and 3±2y in PSP. Included PD patients were 9 male (M), 11 female (F), 6 right body-side affected, 14 left-side affected, with mean age of 63±9y and Disease Duration (DD) at scanning of 3±2 years. Fourteen probable MSA, 7 possible MSA patients (10M, 11F, age 64±10y; DD 4±2y), and 13 probable, 4 possible PSP patients (9M, 8F, age 68±8y; DD 2±1y) were included.

2.2.2 *Feature Extraction*

We reimplemented the SSM/PCA method in Matlab based on the description by Spetsieris and Eidelberg [Eidelberg, 2009; Spetsieris et al., 2010; Spetsieris and Eidelberg, 2011; Spetsieris et al., 2009]. First, the FDG-PET images are loaded in a data matrix P_{sv}, and a mask is applied to each subject image in P_{sv} (s [1,...,M] refers to subjects and the column index v refers to voxels) to remove all voxels with intensity value less than 35% of the whole brain volume maximum. Then the subject matrix is log-transformed and doubly centered to create a subject residual profile (SRP) matrix SRP_{sv}. PCA is then applied to the matrix SRP_{sv} to obtain its eigenvectors. These eigenvectors are called Group-Invariant Subprofile (GIS) patterns ($GIS_k, k = 1, 2, ..., M$), and represent characteristic disease-related brain patterns. Furthermore, subject scores are computed as the contribution of each subject image to a disease related pattern GIS_k.

Figure 2.5 illustrates the flow of the program. The main steps of the method are as follows, see [Spetsieris and Eidelberg, 2011].

1. The FDG-PET images are loaded into a data matrix P_{sv} [1] of dimension $M \times N$, where the row index s (1,...,M) refers to subjects, and the column index v refers to voxels. So row s contains the 3D image data for subject s, reformatted in vector form.
2. A mask is applied to each subject image in P_{sv} to reduce low values and noise in the brain volumes. In this study, all voxels with intensity value less than 35% of the whole brain volume maximum were removed to create individual masks. Next a group mask with non-zero values for

1 By expressions like "the matrix P_{sv}" we mean the matrix P with elements P_{sv}.

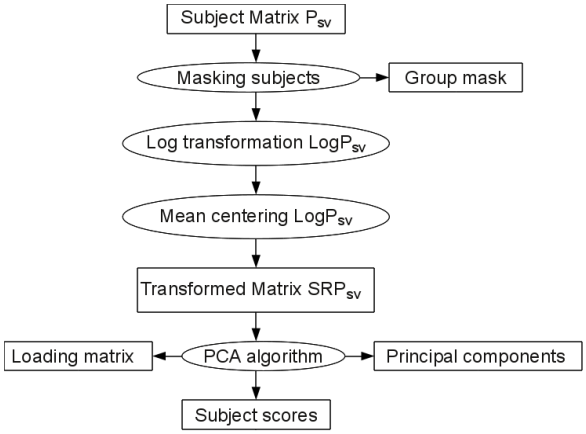


Figure 2.5: Computation flow chart illustrating SSM/PCA analysis. For explanation see text.

all subjects is created by taking the intersection of the individual masks.

3. The subject matrix P_{sv} is log-transformed to obtain the matrix $LogP_{sv}$. This step is necessary in order to remove multiplicative scaling effects.

4. The log-transformed subject matrix is doubly centered. The mean across voxels (local glucose metabolic rate LGMR) for each subject and the mean across subjects (GMP image) is subtracted from the matrix $LogP_{sv}$ to create a *subject residual profile* matrix SRP_{sv}, i.e.,

$$SRP_{sv} = LogP_{sv} - LGMR_s - GMP_v$$

where

$$LGMR_s = mean_{vox}(LogP_{sv})$$
$$GMP_v = mean_{sub}(LogP_{sv}) - mean_{sub}(LGMR_s)$$

Here $mean_{vox}$ is the mean across voxels per subject and $mean_{sub}$ is the mean across subjects per voxel. This double centering is carried out in order to: 1) remove offset differences per subject 2) remove offsets per voxel, i.e., enhance/retain differences between subjects per voxel; removing uninformative overall behaviour.

5. PCA is applied to the matrix SRP_{sv}. The eigenvectors of the PCA analysis are called Group-Invariant Subprofile (GIS) patterns ($GIS_k, k = 1, 2, ..., M$), and represent characteristic brain patterns.

6. Subject scores are computed as the amount of expression of a subject on a disease related pattern GIS_k. The score of subject s for the kth GIS pattern is defined as the inner dot product of the subject's SRP row vector and the kth GIS vector:

$$\text{Score}_{ks} = \text{SRP}_s^T \cdot \text{GIS}_k \tag{2.1}$$

The SSM/PCA method was applied to several data groups (disease group(s) compared to healthy controls) in training set(s) from which disease related patterns (GIS_k) were extracted with positive and negative loadings (voxel weights) [Ma et al., 2009]. The brain images from the training set are weighted onto the patterns to obtain subject scores, which depict how much each subject image contributes to a pattern.

Subject scores as features for classification

Features are usually derived as characteristics of an object such as texture, color, or shape [Westenberg and Roerdink, 2002], which can be computed for each subject (data set) separately. The use of PCA-based subject scores as features deviates significantly from the standard situation through the fact that features now depend on the whole dataset. Also, the number of features is, at least initially, equal to the number of principal components which is equal to the number of data sets. So when a subject is removed or added to the data collection the scores of all the other subjects change as well. Therefore, there is need to redo the SSM/PCA procedure once the dataset changes to obtain new scores.

2.2.3 *Decision tree classification*

The C4.5 decision tree method [Quinlan, 1996b] is a supervised learning strategy which builds a classifier from a set of training samples with a list of features (or attributes) and a class label. The algorithm splits a set of training samples into subsets such that the data in each of the descending subsets are "purer" than the data in the parent subset (based on the concept of information gain from information theory). Each split is based on an optimal threshold value of a single feature. The result is a tree in which each leaf carries a class name and each interior node specifies a test on a particular feature. The tree constructed in

the training phase of a decision tree classifier can be drawn in an easy to understand graphical representation which shows the successive features and threshold values which the algorithm has used to separate the data set in non-overlapping classes. Once a tree has been obtained from the training samples, it can be used for testing to classify unseen cases where the class label is unknown.

The C4.5 decision tree algorithm [Quinlan, 1993] has been used in many previous studies, ranging from diatom identification [du Buf and Bayer, 2002] to classification of anomalous and normal activities in a computer network to curb intrusions [Muniyandi et al., 2012]. The method has also been applied to improve accuracy in multi-class classification problems. For example, Polat and Güneş [2009] applied a novel hybrid classification system based on the C4.5 decision tree classifier and a one-against-all approach, obtaining promising results. In addition, Ture et al. [2009] analysed several decision tree methods (CHAID, CART, QUEST, C4.5, and ID3) together with Kaplan-Meier estimates to investigate their predictive power of recurrence-free survival in breast cancer patients, they report that C4.5 performed slightly better than other methods. In summary, decision trees are considered to be powerful for classification and are easy to interpret by humans. Not only are they simple and effective, but they also work well with large datasets [Perner, 2001].

DECISION TREE CLASSIFICATION OF PARKINSONIAN SYNDROMES Using the C4.5 machine learning algorithm, we trained classifiers on subject scores of extracted patterns for healthy subjects and subjects with known types of neurodegenerative disease. The result is a pruned decision tree showing classified subject images. The goal of pruning is to obtain a tree that does not overfit cases. Note that it would be possible to obtain 100% correct classification in the training phase by using a less stringent pruning strategy. However, this would come at the expense of generalization power on unseen cases.

In contrast to applications of the SSM/PCA method which make a pre-selection of principal components (GIS vectors) on which the classification will be based, the C4.5 algorithm uses all principal components and the corresponding subject scores as input. The algorithm itself determines which principal components are most discriminative to separate the data set

into classes. More discriminative components appear higher in the decision tree, i.e., closer to the root; refer to Figure 2.6 for an example, where the subject score SSPC5 is the most discriminative feature.

In order to apply the C4.5 classifier to unseen cases, the required subject scores for testing are first computed by projecting the SRP of the new subject on the GIS profiles of the training set according to Eq. (2.1). The computation of the SRP for the unseen case involves centering along the subject dimension, i.e., subtracting the GMP (group mean profile). The assumption is that this GMP can be obtained from the reference group only, i.e., the group used for training the classifier; see the discussion in Spetsieris et al. [2009], [p. 1244].

2.2.4 *Other Classifiers*

We also applied a number of other classifiers: nearest neighbors; linear classifiers: linear discriminant analysis and support vector machines; random forests, which is an extension of decision trees; classification and regression trees (CART) for predicting real/continuous variables; and naive Bayes, a probabilistic classifier. Linear classifiers in particular are simple to implement. They are known to work better in situations where the data is uniformly distributed with equal covariance.

NEAREST NEIGHBORS (NN) NN is a classification method which assigns a class to a new data point based on the class of the nearest training data point(s). In the K-NN (K-Nearest Neighbors) method, distances to the neighbors are computed first. Then, a new data point receives the majority label of the K nearest data points.

LINEAR DISCRIMINANT ANALYSIS (LDA) LDA, like PCA, is used for data classification and dimensionality reduction. This classifier maximizes the between-class variance and minimizes the within-class variance to ensure a clear separation in datasets. Accordingly, the training data are first transformed, then the data in the transformed space are classified as belonging to a class which minimizes the Euclidean distance of its mean to the transformed data [Fukunaga, 1990].

SUPPORT VECTOR MACHINE (SVM) SVM performs classification by generating an optimal decision boundary in the form of a hyperplane which separates different classes of data points in the feature space. The decision boundary should maximize the distance between the hyperplane and support vectors called the margin [Duda et al., 2000].

RANDOM FORESTS Random Forests is a machine learning method for classification of objects based on a majority vote of a multitude of decision trees. This method combines bagging (random selection of cases) and random selection of features (at each node) during the training phase. Also, the trees are not pruned.

CLASSIFICATION AND REGRESSION TREES (CART) CART, just like C4.5, is a decision tree learning method. However, in addition to using decision trees as predictors, CART includes regression trees for predicting continuous variables.

NAIVE BAYES This is a method that classifies data points based on their likelihood and the prior probabilities of occurrences of known classes. The final classification is achieved by combining the prior and the likelihood to form a posterior probability using Bayes' rule. Overall, the new data will belong to a class which maximizes the posterior probability.

2.3 RESULTS AND DISCUSSION

2.3.1 *Results for decision tree classifiers*

Decision tree classifiers were trained by applying the C4.5 algorithm to individual (each disease group versus healthy controls) and combined datasets of PD, PSP, MSA patients and healthy controls (HC) with known class labels, as listed in Section 2.2.1. For the individual datasets, we were interested in identifying features which best separate two groups (i.e., a disease group from healthy controls). For the combined datasets we compared all the groups, that is, PD, MSA, PSP, and HC to each other to obtain feature(s) which can separate the four groups. Tree pruning was carried out by using the default values of the C4.5 algorithm [Quinlan, 1993].

2.3.1.1 *Building classifiers for individual datasets*

Decision tree classifiers were built in the training phase from the individual datasets (PD, PSP, MSA) compared to the HC group of 18 subjects.

PD GROUP The decision tree built from the PD-HC dataset (18 healthy and 20 PD subjects) is illustrated in Figure 2.6. The subject scores derived from 38 principal components (GIS vectors) are the attributes on which decisions are made. They are represented as oval-shaped interior nodes in the tree. Next to the arrows the threshold values are shown that were used to split the dataset. Likewise, the leaf nodes, represented as rectangles, show the final class or decision made at that level of the tree (for example, PD or HC in Figure 2.6). Red and blue colors are used to indicate cases labeled as PD and healthy, respectively. The numbers between brackets in the rectangles show the total number of cases classified at that leaf. Additionally, the number after the slash (if present) represents the number of misclassified cases at that leaf.

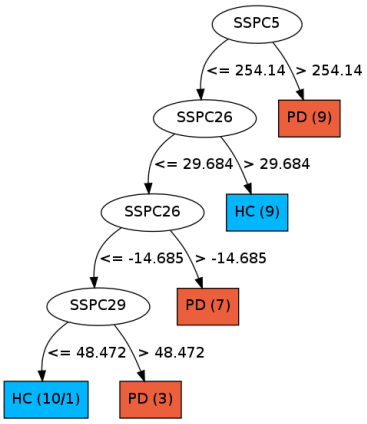


Figure 2.6: The decision tree built from the PD-HC dataset. Oval-shaped interior nodes: features (subject scores) used to split the data. Threshold values are shown next to the arrows. Rectangular leaf nodes: the final class labels (red=PD, blue=HC).

As can be seen in Figure 2.6, the classifier chooses the subject score based on component 5 (SSPC5) to make the first split. In the right subtree, nine PD subjects > 254.14 are identified. The classifier goes on to test the rest of the subjects based on

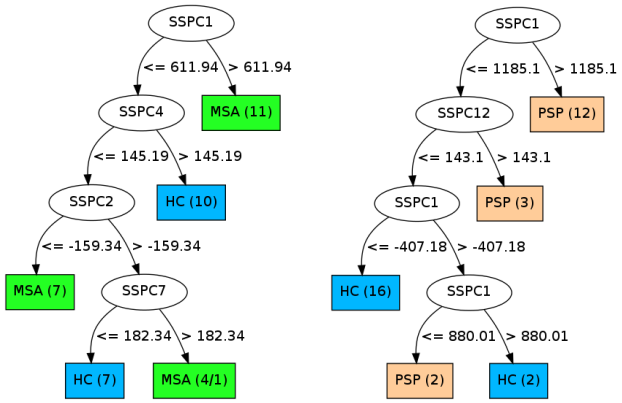


Figure 2.7: The decision trees built from the MSA-HC (left) and PSP-HC (right) datasets. For details, refer to Fig. 2.6.

component 26, where nine subjects (subject score > 29.684) are identified as HC; etc. Only one PD subject is misclassified as HC, as can be seen in Figure 2.6 in the lower-left rectangle.

MSA GROUP The decision tree built from the MSA-HC dataset (18 healthy and 21 MSA subjects) is illustrated in Figure 2.7 (left). The attributes are subject scores derived from 39 principal components. Again, one HC subject is misclassified

PSP GROUP The decision tree built from the PSP-HC dataset (18 healthy and 17 PSP subjects) is illustrated in Figure 2.7 (right). The attributes are subject scores derived from 35 principal components.

2.3.1.2 *Building classifiers on combined datasets*

We also applied the C4.5 classification algorithm to the combined datasets consisting of all four groups. Therefore, the dataset consisted of 76 subjects, 18 HC, 20 PD, 21 MSA and 17 PSP. Subject scores were obtained by applying the SSM/PCA method to the combined group. The resulting decision tree is shown in Figure 2.8. Three PSP subjects are classified erroneously, two as PD and one as MSA.

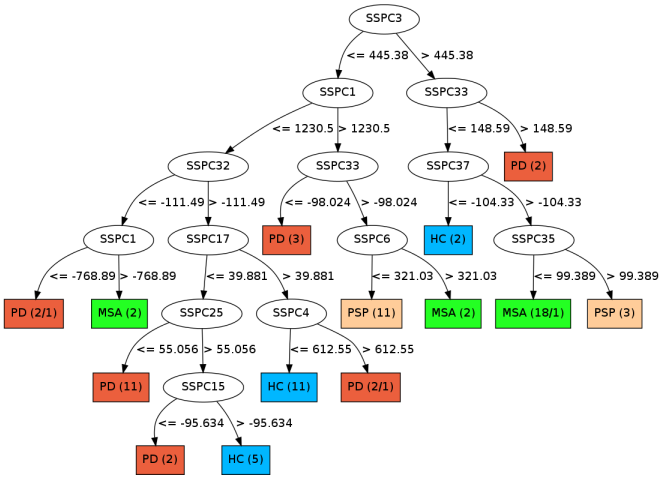


Figure 2.8: The decision tree built from the combined PD-PSP-MSA-HC dataset.

2.3.1.3 *Leave-one-out cross validation*

In leave-one-out cross-validation (LOOCV), a single observation from the original dataset is used as the validation set (also known as test set) and the remaining observations form the training set. This procedure is repeated N times where each observation is used once as a validation set.

The LOOCV method was applied to individual and combined datasets, i.e., PD-HC, MSA-HC, PSP-HC, and the combined dataset PD-MSA-PSP-HC to estimate classifier performance on unseen cases. Here performance is defined as the percentage of correct classifications over the N repetitions. To ensure that attributes of the training set, and thus the trained classifier, are independent of the validation sample, the test subject was removed from the initial dataset before applying the SSM/PCA method to the training set (with $N-1$ samples) for obtaining the subject scores needed to train the C4.5 decision tree classifier. The classifier was then used to determine the label for the test subject. This procedure was applied for each of the N subjects in the original dataset. Table 2.1 shows the classifier performance.

As seen in Table 2.1, the C4.5 classifier performs highest with the PSP group at 80% and lowest with the PD group at 47.4%. The feature at the root of a decision tree is most significant in classification, since it has the highest information gain (see

Table 2.1: Classifier performance for the different data sets (patients vs healthy controls, number of cases in brackets) in the LOOCV, without feature pre-selection. The column Perf.(%) indicates the percentage of subject cases correctly classified per group, Sensitivity (%) the percentage of correctly classified patients and Specificity (%) the percentage of correctly classified healthy controls.

Feature set(size)	Perf. (%)	Sensitivity (%)	Specificity (%)
PD-HC (38)	47.4	45	50
MSA-HC (39)	71.8	61.9	83.3
PSP-HC (35)	80.0	82.4	77.8

Section 2.2.3). As seen in Figure 2.7, feature 1 (i.e., the subject score on principal component 1) is chosen by the classifier in making a first separation between healthy and PSP/MSA subjects. Moreover, we observed that for the PSP-HC group feature 1 occurs as the root for all LOOCV trees. This behaviour is strongly linked to the high performance for the PSP group, since the classifier is utilizing the relevant feature(s) for the separation of the groups.

The MSA-HC dataset has the second best performance and we observed that the feature at the root of the MSA-HC tree in Figure 2.7 (left) also appears as root in 32 out of 39 trees in LOOCV. On the contrary, for the PD group, different features were chosen by the classifier as root nodes of the different LOOCV trees. Apparently, the different features contain only weakly relevant information to separate the healthy group from the PD group. In this case, application of the decision tree method with all features included leads to a form of over-fitting. We attribute this to the fact that the PD group is quite similar to the HC group, at least with respect to the features we have measured. The early PD group might contain other disease sub-types which need to be identified.

For the combined dataset (see Figure 2.8), feature 3 occurs as the root node, so is the best at separating the four groups (HC, PD, MSA, and PSP). Furthermore, the same feature occurs as the root node in 63 out of 76 LOOCV trees, implying consistency of the classifier. However, the performance for the combined group is low, i.e., 53.9% (the number of correctly classified healthy controls, PD, PSP, and MSA subjects is equal

to 55.6%, 35%, 58.5%, and 66.7%, respectively). Our explanation is that the number of subjects per class is quite low given the large variability in each group. In addition, the combined group is not well balanced in view of a relatively small size of the healthy subject group versus the combination of the three disease groups.

PERMUTATION TEST In order to determine the significance of the performance results we ran a permutation test on the PD-HC, MSA-HC, and PSP-HC groups [Al-Rawi and Cunha, 2012; Golland and Fischl, 2003]. The steps of the procedure are:

1. for each group, perform a LOOCV on the original subject labels to obtain a performance P_O;
2. repeatedly permute the labels and then do a LOOCV to obtain performances P_i for $i = 1, \ldots, N_{perm}$ (we used $N_{perm} = 100$)
3. compute the p-value as the total number of all P_i greater or equal to P_O, divided by N_{perm}.

If $p < 0.05$ the original LOOCV result is considered to be statistically significant.

The results of the permutation test were as follows. For the PSP-HC group: $p = 0.00$; for the MSA-HC group: $p = 0.01$; for the PD-HC group: $p = 0.62$. So we can conclude that for the PSP-HC and MSA-HC groups the performance results are significant. However, for the PD-HC group this is not the case. This is consistent with the lack of robustness of the LOOCV trees we already noted above. The healthy and PD group are very similar and hard to separate, given the small number of datasets.

2.3.1.4 *Pre-selection of features*

In the hope to improve the classifier performance, we varied the number of features used to build the classifier in the LOOCV. This was done in two different ways: (i) by choosing the subject scores of the n best principal components according to the Akaike Information Criterion (AIC) [Akaike, 1974]; (ii) by choosing the first n principal components arranged in order of highest to lowest amount of variance accounted for. The classifier performance at the varying numbers of features is shown in Table 2.2.

Table 2.2: Classifier performance with pre-selection of features (patients vs healthy controls, number of cases in brackets). The percentage of principal components arranged in order of highest to lowest variance accounted for, and best number of PCs according to AIC. Highest performances in bold.

	In order of amount of variance					According to AIC		
% / no of PCs	3%	5%	50%	70%	100%	1	3	5
PD-HC (38)	55.3	**63.2**	57.9	**63.2**	47.4	**63.2**	50	47.4
MSA-HC (39)	71.8	**74.4**	69.2	71.8	71.8	66.7	69.2	**74.4**
PSP-HC (35)	**82.9**	80	77.1	77.1	80	**82.9**	80	80

As shown in Table 2.2, the performance for the PD group improves from 47.4% to 63.2% when the number of features is reduced from 100% to 70% and 5%. Also the performance improves when only one best feature according to AIC is used to build the classifier. Likewise the performance for the MSA and PSP groups improve from 71.8% to 74.4% and 80% to 82.9%, respectively, when the number of features are reduced. Notable is that the number of features at which distinct groups perform best may differ. Specifically, when using the AIC for pre-selection, not always one feature is good enough to separate groups. This can be seen for the MSA group where five best features were required to obtain the best performance. Overall, pre-selection/reduction of features to include relevant features can boost classifier performance.

2.3.1.5 *Disease groups versus each other*

Disease groups were compared to each other in a binary classification. That is to say, the PD group of 20 subjects versus the MSA group of 21 subjects, PD group of 20 versus PSP group of 17 and MSA group of 20 vs PSP group of 17.

As seen in Table 2.3, PD vs MSA has the highest performance with a relatively high sensitivity and specificity, consequently PD can be separated rather well from MSA. For the PD vs PSP and MSA vs PSP groups the performance is slightly lower. The performance for all groups slightly increases when features are reduced to only 5 according to AIC. In spite of the high performance of the PSP group versus the healthy group as seen in Table 2.1, PSP performs relatively low when compared to the other disease groups (PD and MSA). Apparently, the PSP

Table 2.3: Performance for binary classification of disease groups in the LOOCV. The number of cases per group are in brackets. The column Perf. indicates the percentage of subject cases correctly classified (all features included), Sensitivity the percentage of correctly classified first disease group, Specificity the percentage of correctly classified second disease group, and Perf. (AIC-5) the performance when features are reduced to the best 5 PCs according to AIC.

Group	Perf. (%)	Sensitivity	Specificity	Perf. (AIC-5) (%)
PD vs MSA (41)	73.2	70	76.2	78
PD vs PSP (37)	67.6	80	52.9	70.3
MSA vs PSP (38)	68.4	76.2	58.8	71.1

features look more like those of PD or MSA patients than those of healthy controls.

2.3.1.6 *Combined disease groups*

Our main interest is to distinguish the Parkinsonian syndromes from each other. Therefore, we combined all disease groups (i.e., PD, PSP, and MSA) without the healthy controls in a decision tree multi-classification and applied LOOCV (at 100% features used). The performance of the classifier is 65.5%, with 75% correctly classified PD subjects, 47.1% correctly classified PSP subjects, and 71.4% correctly classified MSA subjects. Altogether the PSP group has the lowest number of correctly classified subjects, in agreement with the previous observation that it contains similarities to PD and MSA. Figure 2.9 shows the decision tree diagram obtained after training the classifier with all features. Only one PD subject is misclassified as PSP.

VARYING THE NUMBER OF FEATURES FOR CLASSIFICATION

Several LOOCV experiments were carried out while varying the number of features used to build the classifier. The highest performance was achieved when including 25% of all features. Results for 100, 50, and 25% of all features are shown in Table 2.4.

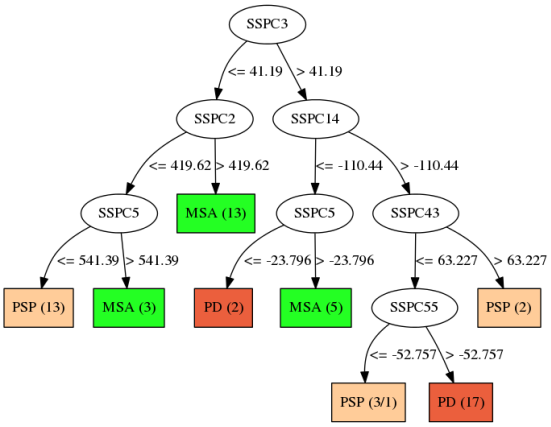


Figure 2.9: The decision tree built from the disease groups compared to each other i.e., PD-PSP-MSA dataset.

Table 2.4: Performance for binary classification of disease groups (number of cases in brackets) in the LOOCV with feature pre-selection. The columns Feat. and Perf. indicate the percentage of features used and the corresponding performance. The remaining columns show confusion matrices and class accuracies. The number of subjects correctly classified for each class is in bold.

Feat. %	Perf. %	Class	PD (20)	PSP (17)	MSA (21)
100	65.5	PD	**15**	5	3
		PSP	4	**8**	3
		MSA	1	4	**15**
		accuracy	75	47.1	71.4
50	67.2	PD	**15**	5	2
		PSP	4	**9**	4
		MSA	1	3	**15**
		accuracy	75	52.9	71.4
25	69	PD	**15**	5	2
		PSP	4	**9**	3
		MSA	1	3	**16**
		accuracy	75	52.9	76.2

2.3.2 *Results for other classifiers*

We used 'scikit-learn' [Pedregosa et al., 2011], a software package that includes a variety of machine learning algorithms, to

obtain classification results for a number of other classifiers. The classifiers used were described in Section 2.2.4. In principle, we should test on subject scores obtained from the leave-one-out method before applying the SSM/PCA method. However, this would lead to a very time-consuming procedure. Since our goal is to obtain an impression of the improvements possible by using other classifiers, we instead applied LOOCV on subject scores obtained from applying the SSM/PCA method to the whole training set (all subjects included).

Performances for the PD, MSA, and PSP groups vs healthy controls are shown in Table 2.5. No pre-selection of features was applied.

Table 2.5: The LOOCV Performance for various types of classifier. Features used were the subject scores obtained after applying the SSM/PCA method on all subjects included in the datasets. (*) Note that for LDA only 90% of the features were considered because of the classifier's restrictions while constructing the covariance matrix. For easy reference, the feature pre-selection results for C4.5 already presented in Table 2.2 are included.

Dataset	PD-HC	MSA-HC	PSP-HC
Nearest Neighbors	76.3	76.9	80.0
Linear SVM	78.9	92.3	88.6
Random Forest	63.2	61.5	71.4
Naive Bayes	65.8	71.8	71.4
LDA (*)	50.0	61.5	65.7
CART	57.9	53.8	85.7
C4.5	63.2	74.4	82.9

2.3.3 *Discussion*

The LOOCV performance as shown in Table 2.5 is highest for the SVM and NN classifiers. These classifiers perform better than C4.5, especially for the PD-HC group. We attribute this to the fact that SVM and NN only have one decision boundary. On the other hand, C4.5 has several decision boundaries, one for each internal node of the decision tree. Thus a subject is tested more than once and may become vulnerable to misclassification in the case where the features depict noise or are irrelevant.

CART is quite similar to C4.5; for the PD and PSP groups it has a higher performance, but for MSA it is considerably lower.

Decision tree methods are faced with the problem of overfitting, which causes all training cases to be correctly classified but with limited generalizability. That is, the learned tree tends to be so perfect that it is prone to misclassify unseen cases. Also, providing many features to the decision tree inducer can cause a low performance due to irrelevant and redundant features, especially when the number of subjects is relatively small. Moreover it has been observed that C4.5's feature selection strategy is not optimal, so having irrelevant and correlated features can degrade the performance of the classifier [Perner, 2001]. In addition, the C4.5 classifier has been reported to perform lower when it comes to continuous attributes, which is the case in our study (as subject scores are continuous) [Quinlan, 1996a]. However, with pre-selection of features and pruning decision trees after construction, these problems can be reduced. Indeed, we found an increase in performance, especially for the PD-HC group (see Table 2.2).

When the number of subjects in the training set is large enough, the decision tree classifier will be capable to perform sub-type classification of parkinsonian syndromes. Another important advantage of the decision tree method over most other methods is that it provides an intuitive way to get insight in the behavior of the classification algorithm to physicians. Drawings of decision trees are human understandable, and the way a decision tree algorithm takes repeated decisions with respect to multiple criteria is close to the way humans carry out multi-criteria decision making. Likewise, the significance of a particular feature is recognizable from the level in which the corresponding node appears in the constructed tree. Therefore, we have the opportunity to use human intelligence in the decision tree method to select those features (i.e., the corresponding disease related patterns) that best distinguish between healthy subjects and patients.

2.4 CONCLUSIONS

Using the SSM/PCA method, Group-Invariant Subprofile (GIS) patterns were extracted from FDG-PET data of patients with three distinct groups of syndromes, i.e., Parkinson's disease (PD), multiple system atrophy (MSA), and progressive supranu-

clear palsy (PSP), always compared to a healthy control (HC) group. The subject scores corresponding to these patterns served as the feature set for the C4.5 decision tree classification algorithm. Classifiers were constructed for future prediction of unseen subject images. Validation of classifiers to ensure optimal results was performed using the leave-one-out cross-validation (LOOCV) method. A permutation test was performed to assess the statistical significance of the results.

We also compared the C4.5 classifier to various other classification algorithms, i.e., Nearest Neighbors, Linear SVM, Random Forest, Naive Bayes, LDA, and CART. Of all classifiers, the performance of Nearest Neighbors and Linear SVM was highest. We found that most classifiers perform relatively well for the PSP-HC and MSA-HC groups, but less well for the PD-HC group. This may be closely linked to the fact that the FDG-PET activation pattern of (early stage) PD patients is close to that of normal subjects, whereas there is one distinctive feature which is present in MSA (low uptake in putamen) and PSP (low frontal uptake), respectively, and absent in controls.

In clinical practice, the main problem is not so much to distinguish patients with parkinsonian syndromes from healthy controls, but to distinguish between the different parkinsonian diease types. For this reason, we also compared disease groups to each other in a binary classification, with promising results: in this case classifier performance was significantly higher, also when the PD group was involved. In a recent study, Garraux et al. [2013] used Relevance Vector Machine (RVM) to classify 120 parkinsonian patients on the basis of either binary classification (a single class of 3 atypical parkinsonian syndromes [APS] versus PD), or multiple classification (PD and the 3 APS separately versus each other). The performance achieved in the study of Garraux *et al.* was higher than in ours. Note, however, that they had a larger dataset and incorporated bootstrap aggregation (bagging) to boost the performance. We plan to incorporate bagging in future work to improve classifier performance.

To achieve high-quality biomarker identification, one needs to accumulate large numbers of patient data in several phases of disease progression. This is what we are currently pursuing in the GLIMPS project [Teune et al., 2012], which aims at establishing a national database of FDG-PET scans in the Netherlands. Additionally, data could be generated from other

imaging modalities such as (f)MRI, ASL, and DTI, to enable the collection of a broad set of brain features needed for distinguishing the different disease types.

3

COMPARISON OF DECISION TREE AND STEPWISE REGRESSION METHODS IN CLASSIFICATION OF FDG-PET BRAIN DATA USING SSM/PCA FEATURES

ABSTRACT:

Objective: To compare the stepwise regression (SR) method and the decision tree (DT) method for classification of parkinsonian syndromes.

Method: We applied the scaled subprofile model/principal component analysis (SSM/PCA) method to FDG-PET brain image data to obtain covariance patterns and the corresponding subject scores. The subject scores were inputs to the C4.5 decision tree algorithm to classify the subject brain images. For the SR method, the scatter plots and receiver operating characteristic (ROC) curves show the subject classifications. We then compare the decision tree classifier results with those of the SR method.

Results: We found out that the SR method performs slightly better than the DT. We attribute this to the fact that the SR method uses a linear combination of the best features to form one robust feature, unlike the DT method. However, when the same robust feature is used as input to the DT classifier, the performance is as high as that of the SR method.

Conclusion: Even though the SR method performs better than the DT method, including the SR procedure in the DT classification yields a better performance. Additionally, the decision tree approach is more suitable for human interpretation and exploration than the SR method.

KEYWORDS: *Parkinsonian syndromes, FDG-PET data, scaled subprofile model, principal component analysis, decision tree classification, stepwise regression.*

3.1 INTRODUCTION

Parkinsonian syndromes like other neurodegenerative diseases are not easy to diagnose and distinguish at an early stage [Spetsieris et al., 2009; Wu et al., 2013]. With the intention to classify these syndromes, the scaled subprofile model/principal component analysis (SSM/PCA) method as explained by Moeller et al. [1987] is used to extract disease-related metabolic brain patterns in the form of principal component images from subject brain images. Then individual subject images are projected onto the patterns to obtain their corresponding scores. These scores depict the network expression of individual subjects on the pattern [Fukunda et al., 2001].

The SSM/PCA method has been used in several studies to extract disease-related patterns from imaging data. In Moeller et al. [1996], the SSM method is applied to regional metabolic rates for glucose data to identify specific age-related disease profiles. Similarly, in Spetsieris et al. [2009] the SSM/PCA method is used to derive disease-related spatial covariance patterns which are represented as spatial weighted images. In the study by Spetsieris and Eidelberg [2011] the methodological questions that arise regarding the use of the SSM method are addressed. In addition, the SSM/PCA method together with several versions of the Statistical Parametric Mapping (SPM) software were applied by Peng et al. [2014] to obtain disease-specific patterns. Therefore, from the aforementioned studies we can say that the SSM/PCA method application is quite broad and effective at identifying brain patterns. These patterns are promising as biomarkers for predicting Parkinsonian disorders and neurodegenerative diseases in general.

This chapter presents a comparison between the stepwise regression (SR) method [Teune et al., 2013] and the decision tree (DT) method in the classification of parkinsonian syndromes, following previous work [Mudali et al., 2015]. In both methods we apply the SSM/PCA method to the brain data to obtain subject scores as features. Specifically, we use the C4.5 machine learning algorithm in this study to build the DT classifiers [Quinlan, 1993, 1996b; Polat and Güneş, 2009]. The SR method uses a mechanism of choosing one or a few models (here known as components) from a larger set of models [Johnsson, 1992; Thompson, 1995]. Further, the components are chosen based on how well they separate subject image groups using the Akaike information criterion (AIC) [Akaike, 1974].

There are three approaches we use in this study:

1. the stepwise regression (SR) method
2. decision tree classification with all features, and a reduced set of features, respectively
3. decision tree classification using the set of features obtained from the SR procedure.

With the SR method, one feature (subject z-score) is determined from a combination of components, while in the DT method several features (subject scores) are determined from individual components. In approach 3 we combine the SR procedure and decision tree method in two different ways. In the first approach, the best features obtained by the stepwise procedure

are used as features for decision tree classification, that is, without linearly combining them. In the second approach, we use the exact same subject z-score (that is, a linear combination of best features) as obtained by the SR method (stepwise plus logistic regression procedure) and use it as a single feature for decision tree classification.

3.2 METHOD

3.2.1 *Data acquisition and feature extraction*

We used fluorodeoxyglucose positron emission tomography (FDG-PET) brain scans as described in the previous studies by Teune et al. [2010, 2013]. The data set includes a total of 76 subject brain images, namely: 18 healthy controls (HC), 20 Parkinson's disease (PD), 21 multi-system atrophy (MSA), and 17 progressive supra-nuclear palsy (PSP). An implementation of the SSM/PCA method developed in Matlab was used following the procedure as described by Eidelberg [2009]; Spetsieris et al. [2009, 2010]; Spetsieris and Eidelberg [2011].

The SSM/PCA method was applied to the FDG-PET data to obtain principal components (PCs) onto which original images were projected to obtain their weights on the PCs, known as subject scores. Thereafter, we used the subject scores as features for the decision tree method and the stepwise regression procedure to differentiate among the parkinsonian syndromes.

3.2.2 *Classification*

3.2.2.1 *Stepwise regression method*

Following Teune et al. [2013], the SR procedure is used to obtain a linear combination of PCs (combined pattern) that best discriminates groups. The SR method is as follows:

- The principal components that make up 50% of the variance are considered in the stepwise regression procedure. This procedure retains only those which best separate groups according to Akaike's information criterion (AIC) [Akaike, 1974].

- By fitting the subject scores corresponding to the retained PCs to a logistic regression model, scaling factors for all PCs are obtained. The combined pattern is a sum of PCs weighted by the scaling factors. Then the subject score on the combined pattern is determined by adding the retained subject scores multiplied by their corresponding scaling parameters.
- Z-scores are calculated and displayed on scatter plots and receiver operating characteristic (ROC) curves are determined. Then a subject is classified according to the z-score cut-off value, which corresponds to the z-score where the sum of sensitivity and specificity is maximised. A subject is diagnosed as a patient if the z-score value is higher than the cut-off value and as a healthy control if it is lower than the cut-off value.

LEAVE ONE OUT CROSS VALIDATION (LOOCV) When using the SSM/PCA-SR method, one subject (for testing) is removed from the training set at a time and the SSM/PCA method is applied to the remainder of the subjects. The stepwise regression procedure is followed to create a combined pattern. The left-out subject scores on the PCs that form the combined pattern are multiplied by the scaling parameters to obtain a single subject score on the combined pattern. Each subject score is transformed into a z-score which then becomes the feature used to separate groups.

3.2.2.2 *Decision tree method*

This method builds a classifier from a set of training samples with a list of features and class labels. We used the C4.5 machine learning algorithm by Quinlan [1996b] to train classifiers based on the subject scores as features. As a result, a pruned decision tree showing classified subject images is generated. Pruning helps to obtain a tree which does not overfit cases. Important to note is that with the decision tree method, the principal components are not combined but instead used individually. Therefore, the DT method uses several features (subject scores on several PCs) unlike the SR method which uses only one feature (z-score).

LEAVE ONE OUT CROSS VALIDATION We placed one subject into a test set and the rest into a training set. Then the SSM/PCA method was applied to the training set to obtain subject scores. These subject scores were used to train the classifier and the test subject was used to test the DT classifier performance. The procedure was repeated for each subject in the dataset. We used AIC in conjunction with the SR procedure to pre-select features for the DT method for improving the classifier performance. Further, we provided the one combined feature from the SR method as input to the DT method.

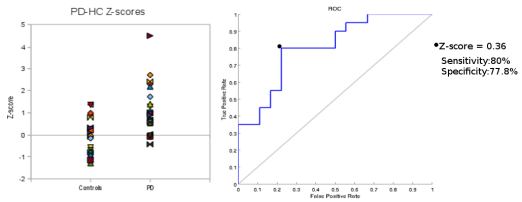
3.3 RESULTS

3.3.1 *Stepwise Regression Procedure*

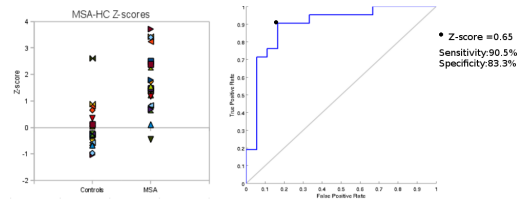
The z-score scatter plots of the combined pattern and the ROC curves are illustrated in Figure 3.10. For the scatter plots, the groups are displayed on the X-axis and the z-scores on the Y-axis. On the ROC curves the bullet (•) represents the cut-point where the difference between true positive rate and false positive rate, or Youden index [Youden, 1950], is maximised. Note that the Youden index is equal to sensitivity+specificity-1 (see appendix 3.A). These results are similar to those in Teune et al. [2013]. The only difference is seen in Figure 3.10(a), where the cut-off is 0.36 instead of 0.45. This can be explained by the fact that at both cut-off points the sensitivity and specificity are the same; in this case 0.36 is chosen being the first z-score value in ascending order.

3.3.2 *Decision tree classifiers for disease groups versus the healthy group*

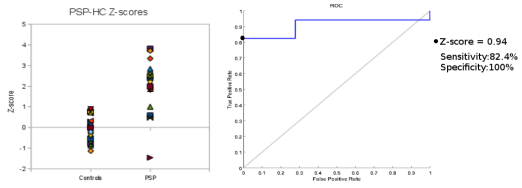
The decision tree classifiers are built from the disease datasets (PD, PSP, MSA) all compared to the healthy control (HC) group of 18 subjects. Figures 3.11 and 3.12 show the decision tree diagrams and corresponding scatter plots. The internal tree nodes are drawn as oval shapes corresponding to the attributes (subject scores) on which decisions are made, with the threshold values for splitting the dataset indicated next to the lines connecting two internal nodes. The actual class labels are represented in the rectangles (leaves), where 1 is the label for



(a)



(b)



(c)

Figure 3.10: Scatter plots and ROC curves for subject z-scores. (a): PD vs HC; (b): MSA vs HC; (c): PSP vs HC.

the disease group (PD, PSP, or MSA) and 0 the label for the healthy group (HC). In addition, the numbers in the brackets of the rectangles show the total number of subjects that are classified at that leaf, with a fraction indicating the number of misclassifications as the denominator.

3.3.2.1 *PD Group*

The output of the decision tree method applied to the PD-HC dataset (18 healthy and 20 PD) is illustrated in Figure 3.11. The attributes are subject scores derived from 38 principal components.

As can be seen in Figure 3.11, the classifier chooses the subject score based on component number 5 (SSPC5) to make the first

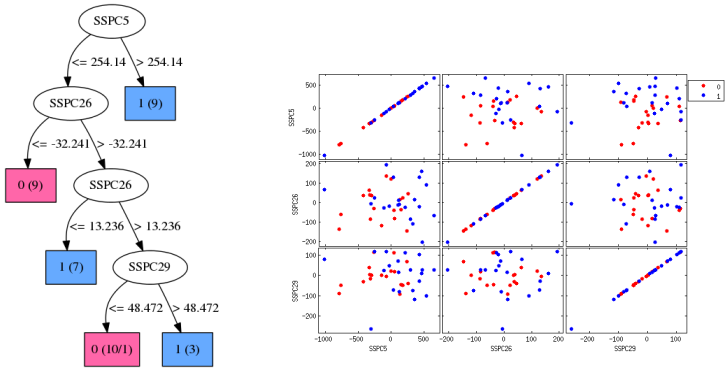


Figure 3.11: The decision tree diagram and the scatter plot showing the distribution of the subject scores of the chosen PCs by the decision tree classifier, without feature pre-selection.

split of the dataset. As a result, nine PD subjects (feature value $>$ 254.14) are identified. The classifier then uses component number 26 to separate the rest of the subjects, where nine subjects (feature value <= -32.241) are identified as HC; etc. Only one PD subject is misclassified as HC. Looking at the scatter plots on the right of Figure 3.11, we can clearly see that for the chosen PCs there is no clear separation between PD and healthy controls.

3.3.2.2 *MSA and PSP Groups*

Figure 3.12 shows the decision trees and the distribution of subject scores displayed on scatter plots for the MSA-HC (18 HC and 21 MSA) and PSP-HC (18 HC and 17 PSP) datasets. The attributes are subject scores derived from 39 and 35 principal components for MSA and PSP, respectively. For the MSA group, one HC subject is misclassified whereas no subject is misclassified for the PSP group. Also, important to note is that for the PSP group the classifier chooses only 2 out of 35 PCs to use, i.e., SSPC1 and SSPC12 as illustrated in the scatter plot of Figure 3.12(b). Moreover, it uses SSPC1 repeatedly to classify the subjects. The C4.5 decision tree inducer can use a feature more than once to classify, as long as it maximizes the information gain.

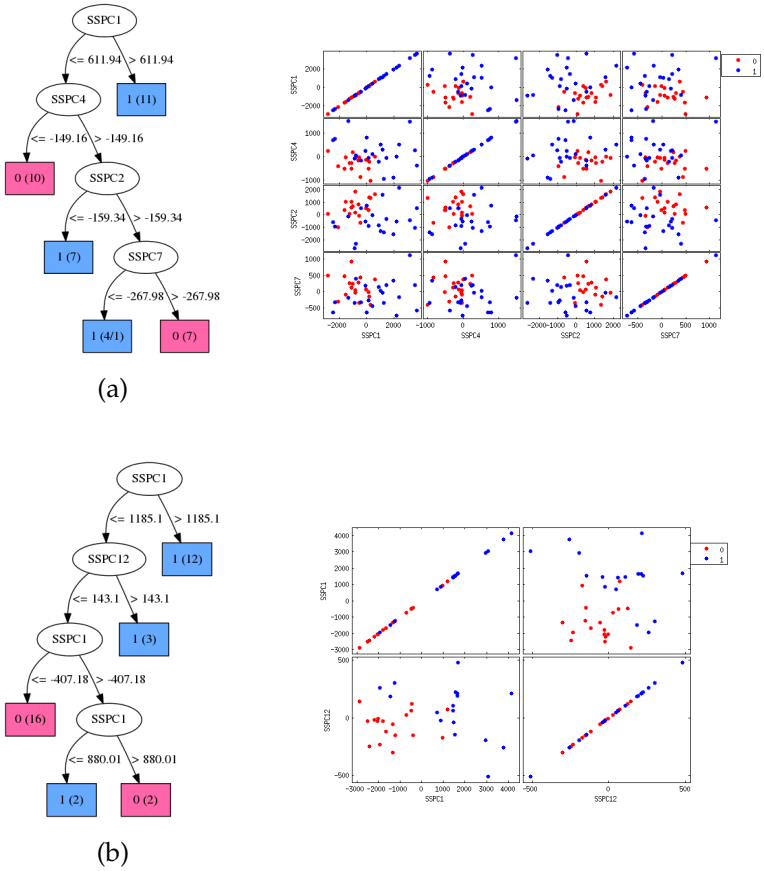


Figure 3.12: Decision tree diagrams and scatter plots showing the distribution of subject scores for the PCs chosen by the classifier. No pre-selection of features. (a): MSA vs HC; (b): PSP vs HC (Note: For the PSP group only two PCs [SSPC1 & SSPC12] were used in the classification).

3.3.3 *Decision trees with reduced number of features*

In Section 3.3.2 we noticed an overlapping distribution of subject scores of the chosen PCs by the classifier with no clear cut between the PD and HC. To improve robustness, we considered to use only the first two components obtained from the PCA process since they depict the highest variance. Figure 3.13(a) is an example of one of the 38 classifiers for the PD vs HC group generated during the LOOCV process, that is the classifier constructed after removing one subject from the training

set, which is thereafter used for testing the left-out subject. For the purpose of comparing with the SR method, we reproduce some of the LOOCV results from the previous study by Mudali et al. [2015], as shown in Table 3.1.

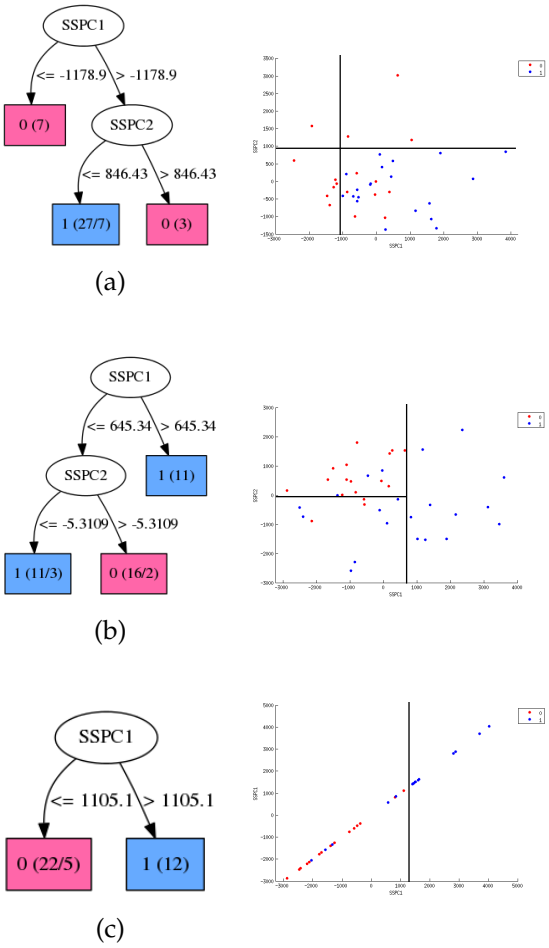


Figure 3.13: The decision tree diagrams and scatter plots showing the distribution of subject scores for the two first features obtained from the LOOCV process. (a): PD vs HC; (b): MSA vs HC; (c): PSP vs HC.

The scatter plot in Figure 3.13(a) shows that there is no clear cut for the classifier to separate the PD and HC groups. This is because the subject scores for both PD and HC are overlapping. As seen from the tree diagram, the classifier chooses one threshold for each of the two given PCs to correctly classify

Table 3.1: Classifier LOOCV performance for reduced number of features, i.e, the first two components according to the highest amount of variance. The column Perf. indicates the percentage of subject cases correctly classified, Sensitivity the percentage of correctly classified patients, and Specificity the percentage of correctly classified healthy controls.

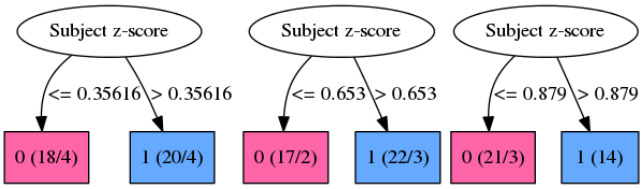
Group	Perf.	Sensitivity	Specificity
PD (38)	63.2	100	22.2
MSA (39)	74.3	83.3	76.2
PSP (35)	80	70.6	88.9

all PD subjects (100% sensitivity), but misclassifies 7/18 HC subjects and the test subject (22.2% specificity). That is to say, the decision boundaries found by the classifier were not successful at efficiently separating the two groups. In this case, even classifiers which use non-axis aligned decision boundaries may not perform well. Accordingly, there is a need to rescale or modify the subject scores (like for the SR method) so that the classifier can find better decision boundaries to efficiently separate the groups.

Unlike the PD-HC group, the MSA-HC group as illustrated in Figure 3.13(b) has a better separation, with the two decision boundaries chosen by the classifier. Only 6/39 subjects overall are misclassified. Important to note is that for the PSP-HC group, the classifier uses only one feature SSPC1 out of the available two features to separate the two groups and 5/17 PSP subjects are misclassified.

3.3.4 *Decision trees with subject z-score on a combined pattern as a single feature*

In the next experiment, the subject z-score determined by the SR method in the study by Teune et al. [2013] is used as a feature for the decision tree classification. This feature is the result of a linear combination of the best PCs according to AIC (for details see Section 3.3.1). Important to note is that we submitted only this single feature (the subject z-score) to the decision tree classifier to separate the patient group from the healthy controls. The results are shown in Figure 3.14 and Table 3.2.



(a) PD vs HC (b) MSA vs HC (c) PSP vs HC

Figure 3.14: The trees obtained after using subject z-score on the combined pattern as a feature for classification.

Table 3.2: Summary of the decision tree classification with the z-score on the combined pattern as a feature.

Group	Perf.	Sensitivity	Specificity
PD (38)	79	80	77.8
MSA (39)	87.2	90.5	83.3
PSP (35)	91.4	82.4	100

In Figure 3.14a the tree chooses a cutoff value 0.36 as the threshold of the single z-score feature to divide the dataset, with 14 out of 18 healthy controls and 16 out of 20 PD subjects correctly classified. These results correspond to the 80% sensitivity and 77.8% specificity at z-score cutoff value of 0.45 as reported in the study by Teune et al. [2013]. That the cutoff values are not identical can be explained as follows. Since the z-scores take a discrete number of values, there can be a small interval of cut-off values which lead to the same sensitivity and specificity (for both the SR and DT method). The decision tree method uses a mechanism called information gain to sort the thresholds in ascending order and then chooses the first threshold. For example, the cut-off interval for the PD group was [0.36,0.45] (with the same sensitivity and specificity), and the decision tree method chose the first which is 0.36. For testing new data samples, a mid-value threshold should be considered to avoid a reduction in specificity.

Interesting is that the DT method produced exactly the same values for sensitivity and specificity as the SR method of Teune et al. [2013], although with small differences in z-score cut-off values. That is to say, at thresholds 0.65 and 0.88 the DT results correspond to the 90.5% sensitivity, 83.3% specificity for the MSA group, and the 82.4% sensitivity, 100% specificity for the

PSP group, respectively, in the study by Teune et al. [2013]. Therefore, with the same single feature (z-score) obtained from a linear combination of the best PCs, the decision tree method is as capable as the SR method (with optimal cut-point value determined from the ROC curve) to obtain high classification performance.

Appendix 3.A illustrates in more detail that maximising the information gain by the DT method and maximising the Youden index [Youden, 1950; Schisterman et al., 2005] in the SR method lead to identical results. We conjecture that this identity holds in more generality, although we do not have a proof at this point.

3.3.5 *Pairwise disease-group comparisons*

In pairwise binary classification we do direct comparisons of each disease against another (that is, excluding the healthy group).

In this experiment, the LOOCV procedure is carried out as usual, but we combine the SR procedure and decision tree method in two different ways. In the first approach, the subject scores of the best components obtained by the SR procedure and AIC (that is, without linearly combining them) are used as features for training the decision tree classifier. In the second approach, we linearly combine these best components to form one pattern and the subject score on the combined pattern is used for training, as in Section 3.3.4. The left-out subject is then tested on the best components (approach 1), or the combined pattern (approach 2).

Table 3.3: Pairwise disease-group comparisons: Classifier LOOCV performance for (1) subject scores on PCs selected by the SR Procedure and AIC; and (2) subject scores on the combined pattern. For each pair of disease groups A and B, Sensitivity is the percentage of correctly classified subjects of group A, and Specificity the percentage of correctly classified subjects of group B.

	Subject scores on individual PCs			Subject scores on combined pattern		
Group	Perf.	Sensitivity	Specificity	Perf.	Sensitivity	Specificity
PD vs MSA (41)	75.6	70	81	90.2	90	90.5
PD vs PSP (37)	70.3	85	52.9	81.1	80	82.4
MSA vs PSP (38)	63.2	66.7	58.8	65.8	61.9	70.6

The PD vs MSA group performs better than the other comparisons for both individual and combined PCs. Similarly, the PD vs PSP group performs well, especially when the PCs are combined. Note that the performance is lowest for the PSP group versus the MSA group. This can be attributed to the fact that PSP and MSA have a quite similar disease pattern [Eckert et al., 2008]. As can be seen, combining of PCs to form one pattern is always better than using individual PCs for the disease pairwise comparisons.

3.4 DISCUSSION

The SR method was found to work better than the DT method, especially when considering all or a few features for the DT method. In most cases the major difference was notable in the performance of the PD vs HC group, which can be attributed to the fact that the PD-related pattern is very similar to the healthy pattern. Additionally, with the PD vs HC comparison, the principal components generated have less variance. Hence, a combination of several best components yields better results, which is exactly what the SR method does.

Furthermore, when the same single z-score feature corresponding to the combined pattern in the SR method is used in the DT classification (see Section 3.3.4), the performance is as high as that of the stepwise regression method [Teune et al., 2013]. The pairwise disease comparisons yielded quite an impressive performance, especially for the PD vs MSA group, when compared to those in Mudali et al. [2015]. Combining the SR procedure with the DT method improved performance in the separation of some disease groups. Therefore, the robust feature obtained using the SR procedure could be used in the DT method to improve classification.

3.5 CONCLUSION

Covariance patterns were extracted from four distinct groups of FDG-PET data using the SSM/PCA method. The subject scores served as the feature set and input to the C4.5 decision tree classification algorithm. Classifiers were constructed from distinct groups for future prediction of new unlabeled subject images. Validation of classifiers was performed using the leave-

one-out method. The decision tree results were compared to the scatter plots and receiver operating characteristic (ROC) curves obtained in the stepwise regression method.

In some instances, the DT results are still not competitive with the SR method. This is because for the decision tree method to maximise classifier performance, it would require several horizontal and vertical decision boundaries to separate the dataset (especially for the PD group) since the subject scores overlap in the feature space. But this could lead to a high generalization error. Hence, it is preferable to combine the features to form one robust feature (subject z-score) which is capable of separating the groups while minimizing the generalization error. In fact, when we included the z-score feature (as used by Teune et al. [2013]) in the DT classification, we obtained identical results for the C4.5 algorithm and the SR method. Therefore, we can improve the DT method by using the linearly combined features obtained by the SR procedure. It would be interesting to find out the performance of a multi-class classification of all parkinsonian syndromes, i.e., PD vs MSA vs PSP using SR feature(s) in the DT classification. Unfortunately, with the SR method in its current form only two groups can be compared.

Nevertheless, given the small size of the current datasets the decision tree method is highly promising. In addition it provides a visual understanding of the classification results and accommodates multi-class classification, as reported in Mudali et al. [2015]. In the long run, we need to devise means of obtaining a more diverse set of features and / or a larger set of training data for the decision tree to perform even better.

3.A APPENDIX: INFORMATION GAIN VERSUS YOUDEN INDEX

In this appendix we consider a data set with healthy and non-healthy cases and compute the optimal split of this data set based on a single attribute according to two different criteria: information gain (as used in decision tree classifiers) and the Youden index. We will illustrate by an example that these two different measures give identical results.

Computing the information gain

Let T be a set of cases, where each case belongs to one of k classes $C_1, C_2, \ldots, C_k$. (e.g., $k = 2$, i.e., healthy and disease.) Let $freq(C_j, T)$ be the number of cases belonging to class C_j.

The *information* of T is:

$$info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} \log_2 \left(\frac{freq(C_j, T)}{|T|} \right) \tag{3.1}$$

When T is split in subsets $T_1, T_2, \ldots, T_n$ by some attribute X which has n outcomes, the *expected information* of T with respect to X is:

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} info(T_i) \tag{3.2}$$

Now consider the complete data set T, with attributes $X_1, X_2, \ldots$. The proportion of healthy cases is p_H, the proportion of disease cases is p_D. Let $info(T)$ be the *information* (entropy) of T. (For a pure set, for example if there are only healthy cases, $info(T) = 0$.) Consider an attribute X and a *split value* V of this attribute. Split the data set T into two subsets T_1 and T_2:

$$\begin{cases} T_1 &= \text{all cases from } T \text{ where } X \leqslant V \\ T_2 &= \text{all cases from } T \text{ where } X > V \end{cases} \tag{3.3}$$

The expected *information* of this partition of T is denoted by $info_X^{(V)}(T)$.

The information gain is: $gain_X^{(V)}(T) = info(T) - info_X^{(V)}(T)$. In order to find the optimal split of the data set, one computes $gain^{(V)}(X)$ for all attributes X and all split values V. Then the attribute X which maximizes $gain_X^{(V)}$ is chosen as the first node of the tree, with V the corresponding split value.

Youden index

For distinguishing between individuals with and without a disease, the *Youden index* is often used [Youden, 1950; Schisterman et al., 2005] as a measure of overall diagnostic effectiveness. This index is defined by $J = TPR - FPR$, with TPR the true

positive rate (fraction of true positives out of all positives), and FPR the false positive rate (1-fraction of true negatives out of all negatives). In other words, J is the maximum vertical distance between the ROC curve and the diagonal or chance line. Note that TPR equals sensitivity and FPR equals 1-specificity, so that J is equal to sensitivity+specificity-1.

Example

Consider now an example data set T with six cases, two healthy (labeled H) and four diseased (labeled D). Let us consider the disease cases as positives and the healthy cases as negatives. We now consider all possible choices for the split point; let us indicate the cases by 0,1,2,...,6. This leads to the seven pictures in Fig. 3.15.

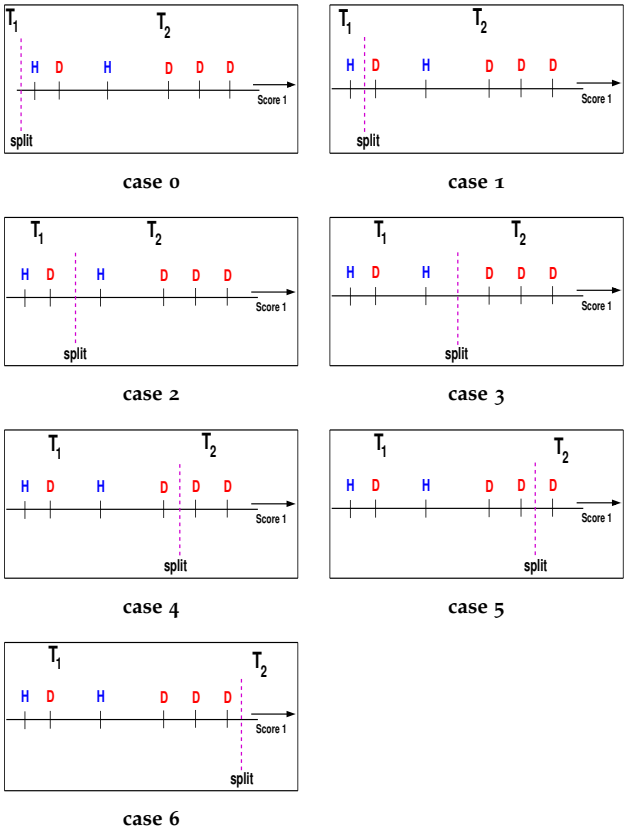


Figure 3.15: All possible cases for the split point.

For all these cases we have computed the Youden index and the information gain $gain_X^{(V)}(T) = info(T) - info_X^{(V)}(T)$, where V refers to the possible cases 0,1,2,...,6 for choosing the split value. Table 3.4 shows the results.

Table 3.4: The Youden index and information gain computed for all the seven cases.

Case No	0	1	2	3	4	5	6
Youden index	0	0.5	0.25	0.75	0.5	0.25	0
Information gain	0	0.32	0.05	0.46	0.25	0.11	0

As can be seen in Table 3.4, Case 3 has both the highest Youden index J and the highest information gain $gain_X^{(V)}(T)$. This illustrates the relationship between the information gain (the mechanism used in the C4.5 decision tree inducer to determine thresholds) and the Youden index used to determine the best cut-off point (best combination of sensitivity and specificity) on the ROC curve, as used in the SR method [Teune et al., 2013].

4

LVQ AND SVM CLASSIFICATION OF FDG-PET BRAIN DATA

ABSTRACT: *We apply Generalized Matrix Learning Vector Quantization (GMLVQ) and Support Vector Machine (SVM) classifiers to fluorodeoxyglucose positron emission tomography (FDG-PET) brain data in the hope to achieve better classification accuracies for parkinsonian syndromes as compared to the decision tree method which was used in previous studies.*

The classifiers are validated using the leave-one-out method. The obtained results show that GMLVQ performs better than the previously studied decision tree (DT) method in the binary classification of group comparisons. Additionally, GMLVQ achieves a superior performance over the DT method regarding multi-class classification. The performance of the considered SVM classifier is comparable with that of GMLVQ. However, in the binary classification, GMLVQ performs better in the separation of Parkinson's disease subjects from healthy controls. On the other hand, SVM achieves higher accuracy than the GMLVQ method in the binary classification of the other parkinsonian syndromes.

KEYWORDS: *Learning Vector Quantization, Support Vector Machine, Parkinsonian syndromes, Classification.*

4.1 INTRODUCTION

Diagnosis of neurodegenerative diseases (NDs), especially at an early stage, is very important to affect proper treatment [Appel et al., 2015], but it is still a challenge [Silverman, 2004]. Nevertheless, some studies report considerable success in differentiating between some of these diseases [Van Laere et al., 2006]. In fact, promising classification performances were obtained for the multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) groups versus the healthy control group in the study [Mudali et al., 2015] where the decision tree (DT) method was used. The same study showed that discriminating the Parkinson's disease (PD) group from healthy controls (HC) on the basis of PET brain scan imaging data remains a challenge. Therefore, in this chapter other classification methods are applied in the hope to improve classification of parkinsonian syndromes, in particular PD, MSA, and PSP. The classification methods used in this study are Generalized

Matrix Learning Vector Quantization (GMLVQ) and Support Vector Machine (SVM).

LVQ is a method which uses prototypes assigned to each class. A new case is classified as belonging to the class of the closest prototype [Kohonen, 1998]. In the training phase, a set of appropriately chosen prototypes is computed from a given set of labeled example data. This training process can be based on a suitable cost function, as for instance in the so-called Generalized LVQ (GLVQ) introduced in [Sato and Yamada, 1996]. The conceptional extension to matrix-based relevance learning was introduced in [Schneider et al., 2009]; simpler feature weighting schemes had been considered earlier in [Hammer and Villmann, 2002]. Relevance learning provides insight into the data in terms of weighting features and combinations of features in the adaptive distance measure. Moreover, GMLVQ allows for the implementation of multi-class classification in a straightforward way.

The Support Vector Machine is a supervised learning method for classifying data by maximizing the margin between the defined classes, see for instance [Burges, 1998; Cristianini and Shawe-Taylor, 2000]. The aim of SVM training is to minimize the classification error while maximizing the gap or margin between the classes by computing an optimally separating hyperplane. The training data points that lie closest to the hyperplane define the so-called support vectors [Cortes and Vapnik, 1995; Zhang, 1999]. This method was originally designed for binary classification but has been extended to multi-class classification, see for instance [Hsu and Lin, 2002] and references therein. Moreover, several studies including [Magnin et al., 2009; Haller et al., 2012] have used SVM to classify neurodegenerative diseases with high accuracy. Other examples of SVM applications like biological data mining are described in [Cristianini and Shawe-Taylor, 2000].

4.2 METHOD

The data used in this study is described in [Teune et al., 2010]. The brain data were obtained from 18 healthy controls (HC), 20 Parkinson's Disease (PD), 17 progressive supranuclear palsy (PSP) and 21 multiple system atrophy (MSA) cases. We apply the scaled subprofile model with principal component analysis (SSM/PCA), based on the methods by Spetsieris et al. [Spet-

sieris et al., 2009], to the datasets to extract features. The method was implemented in Matlab R2014a. The SSM/PCA method [Moeller et al., 1987; Moeller and Strother, 1991; Spetsieris and Eidelberg, 2011] starts by double centering the data matrix and then extracts metabolic brain patterns in the form of principal component images, also known as *group invariant subprofiles*. The original images are projected onto the extracted patterns to determine their weights, which are called *subject scores*. The subject scores then form the features that are input to the classifiers to classify the subject brain images. Because of the application of the PCA method, the computed subject scores are dependent on the whole input dataset, an unusual circumstance in the standard situation. This makes the number of features extracted equal to the number of samples in the dataset.

A leave-one-out cross validation (LOOCV) of the classifiers is performed to predict their performance on new subject cases. For each run, a subject (test sample) is left out, then the SSM/PCA process is performed on the rest of the subjects (training set) to obtain their scores on the principal components. These subject scores are then used to train the GMLVQ and the SVM classifiers. The test subject is projected onto the invariant profiles to obtain its scores on the extracted profiles. Then the test subject scores are used to evaluate the trained classifier. The sensitivity (true positive rate), specificity (true negative rate) and classifier accuracy are determined. Note that the test subject is removed *before* the SSM/PCA process in order to deal with dependencies of the extracted features on both the training and test sets. In addition, the test set receiver operating characteristic (ROC) curve and Nearest Prototype Classifier (NPC) confusion matrix are computed for all the left-out subjects. The area under the curve (AUC) of the ROC curve is a measure of the ability of the features (i.e., subject scores on the principal components) to separate the groups.

For both the SVM and GMLVQ classifiers, we do binary and multi-class classification. The binary classification involves comparing the distinct disease groups (PD, PSP, and MSA) with the healthy control group. The multi-class classification concerns the comparison of all the groups, i.e., HC versus PD versus PSP versus MSA (a total of 76 subjects), as well as only the disease groups, i.e., PD versus PSP versus MSA (a total of 58 subjects). The goal is to determine the class membership

(healthy or diseased) of a new subject of unknown diagnosis and also determine the type of parkinsonian syndrome.

For SVM training and testing, we use the Matlab R2014a functions "fitcsvm" and "predict", respectively, with default parameters and a linear kernel, representing a large margin linear separation in the original feature space. Also, all features are centered at their mean in the dataset and scaled to have unit standard deviation. The "fitcsvm" returns an SVM classifier which can be used for classification of new data samples. It also provides class likelihoods which can be thresholded for an ROC analysis. For the SVM multi-class classification we use the LIBSVM library [Chang and Lin, 2011] with the one-against-one method, since the previously mentioned Matlab functions support only binary classification. The one-against-one method has a shorter training time than the one-against-all, as reported in [Hsu and Lin, 2002].

As for GMLVQ, we employ it in its simplest setting with one prototype $\boldsymbol{w_k}$ per class. A global quadratic distance measure of the form $d(\boldsymbol{w_k}, \boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{w_k})^T \Lambda (\boldsymbol{x} - \boldsymbol{w_k})$ is used to quantify the dissimilarity of an input vector $\boldsymbol{x}$ and the prototypes. The measure is parameterized in terms of the positive semi-definite relevance matrix Λ [Schneider et al., 2009]. Both, prototypes and relevance matrix are optimized in the training process which is guided by a suitable cost function [Schneider et al., 2009]. We employed the gmlvq-toolbox [Biehl, 2015], which performs a batch gradient descent minimization with automated step size control, see [Biehl, 2015] for details. All the results presented here were obtained using the default parameter settings of [Biehl, 2015]. After 100 gradient steps, the training errors and cost function appeared to have converged in all considered classification problems.

It has been shown theoretically and observed in practice frequently that the relevance matrix in GMLVQ displays a strong tendency to become singular [Schneider et al., 2009; Biehl et al., 2015; Bunte et al., 2012]. Generically the relevance matrix is clearly dominated by very few or even a single eigenvector, depending on the complexity of the dataset. This feature of GMLVQ helps to reduce the risk of over-fitting: The effective number of degrees of freedom remains linear in the dimension of the feature vectors, while the number of matrix elements is quadratic. Moreover, GMLVQ provides a low-dimensional

representation of the dataset which can be employed for discriminative visualization, for instance.

4.3 RESULTS

4.3.1 *Generalized Matrix Relevance LVQ (GMLVQ)*

As mentioned earlier, in order to validate the classifiers the training process is repeated with one test subject removed from the training set before applying the SSM/PCA process. This section presents the LOOCV results for the distinct disease groups versus the healthy control group in the binary and multi-class classification. Important to note is that all the features (100%) as extracted from the brain image data using the SSM/PCA method are provided to the GMLVQ classifier. In the tables, sensitivity (%) is the percentage of correctly classified patients, specificity (%) the percentage of correctly classified healthy controls, and AUC is the area under the ROC curve. In addition, the corresponding results are visualized in terms of projections on the leading two eigenvectors of the relevance matrix. This exploits the fact that GMLVQ displays a tendency to yield low-rank matrices which correspond to an intrinsically low-dimensional representation of the feature space [Schneider et al., 2009; Bunte et al., 2012]. Additionally, we include the corresponding plots showing diagonal and off-diagonal matrix elements for one LOOCV iteration as an example illustration.

4.3.1.1 *Binary Classification*

The objective here is to separate the individual disease groups from the healthy control group. The GMLVQ results are shown in Table 4.1.

The results in Table 4.1 are much better than those of the decision tree as reported in [Mudali et al., 2015]. In fact a tremendous improvement can be seen in the PD vs HC group, whose LOOCV performance has increased from 63.2% (decision trees) to 81.6% (GMLVQ). The use of the relevance matrix to weight features according to their relevance appears to boost performance. An illustration is shown in Fig. 4.16 where the training data points are displayed in a feature space of the two leading eigenvectors of the relevance matrix. Observe that the subject scores do not overlap after the GMLVQ classifier

Table 4.1: GMLVQ Classifier performance in LOOCV for the different data sets (patients vs healthy controls, number of cases in brackets). The column Perf.(%) indicates the percentage of subject cases correctly classified per group. Perf. as well as Sensitivity and Specificity correspond to the Nearest Prototype Classifier (NPC).

Feature set(size)	Perf. (%)	Sensitivity (%)	Specificity (%)	AUC
PD-HC (38)	81.6	75	88.9	0.84
MSA-HC (39)	92.3	90.5	94.4	0.99
PSP-HC (35)	88.6	82.4	94.4	0.97

training phase, which corresponds to error-free classification of the training set. Further, the resulting AUC measures (for the different groups) are relatively high. This means that the GMLVQ weighted features are very suitable for separating the groups.

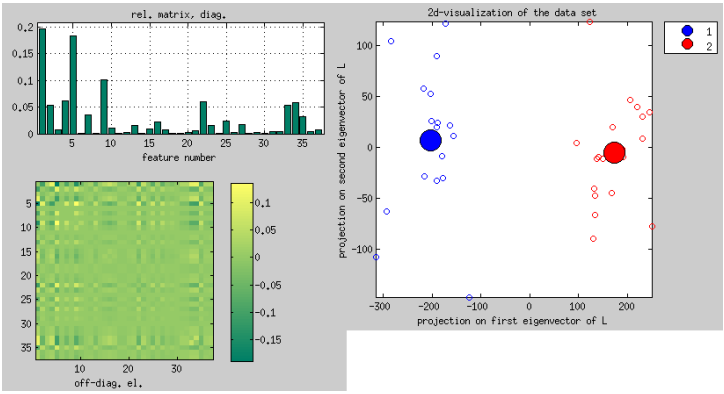


Figure 4.16: Illustrations of the results of a single GMLVQ training process in the LOOCV of the PD vs HC two class-problem, 1 = HC, 2 = disease group. Graphs show diagonal relevances (upper left), and off-diagonal relevance matrix elements (lower left). The visualization of the training data in terms of their projection on the two leading eigenvectors of the relevance matrix is displayed on the right.

As observed in Fig. 4.16, the PD vs HC comparison shows a clear separation between the PD group and the healthy group. Apart from a few outliers, most of the data points cluster around the specific prototypes, i.e., the two bigger circles that

each represent a class. Further, the relevance matrix histogram shows the features and their diagonal weights as used in the classification process. For example, in the PD vs HC group feature 1 was weighted the highest, implying that feature 1 carries relevant information required to separate the two groups. As a matter of fact, the highly weighted feature should be given more attention, i.e., critically analyze the principal component image corresponding to this feature to gain insights from the clinical perspective.

4.3.1.2 *Multi-class classification*

Here we show the results for the LOOCV of the GMLVQ classifier on the multi-class datasets, i.e., the classification of all the four classes, and the three disease classes, respectively. The latter is considered separately, because the main task in clinical practice is to distinguish the three parkinsonian syndromes. Additionally, for the four-class comparison, we include the HC group because we want to build a classifier which can also distinguish a healthy subject from the parkinsonian groups. The results are shown in Tables 4.2 and 4.3 for four-class comparison and three disease groups, respectively. Also included are the scatter plots showing the distribution of training data points in the two-dimensional projection of the feature space in a single run of the training process.

Table 4.2: Four-class problem: The table shows the number of subject images correctly classified for each class in bold and the overall performance in percentage as obtained in the LOOCV.

GMLVQ classification	HC	PD	PSP	MSA
HC(18)	**14**	3	1	0
PD(20)	5	**13**	1	1
PSP(17)	2	2	**11**	2
MSA(21)	0	1	4	**16**
Class accuracy (%)	77.8	65	64.7	76.2
Overall performance (%)	71.1			

FOUR-CLASS COMPARISON. From the results in Table 4.2, we notice that most of the misclassified HC subjects are classi-

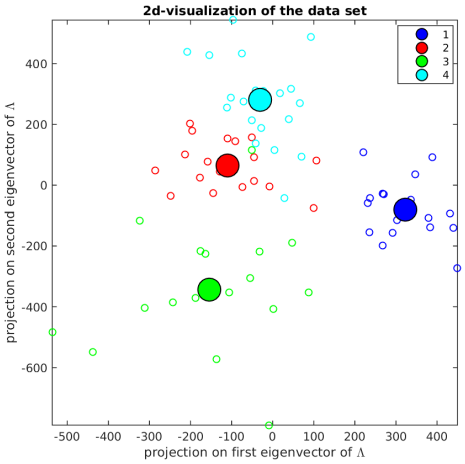
fied as PD and *vice versa*. As already observed in [Mudali et al., 2015], the PD and HC subjects have a closely related metabolic pattern. Likewise, the PSP and MSA groups display a similarity, in view of the fact that four (majority of the misclassification) MSA subjects are misclassified as PSP.

Table 4.3: Three-class problem: The table shows the number of subject images correctly classified for each class in bold with the overall LOOCV performance in percentage.

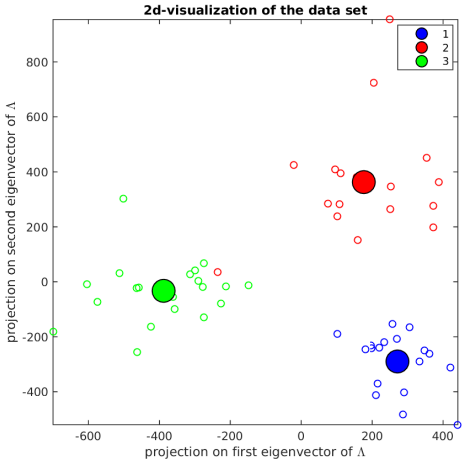
GMLVQ classification	PD	PSP	MSA
PD(20))	**19**	0	1
PSP(17)	2	**12**	3
MSA(21)	2	3	**16**
Class accuracy (%)	95	70.6	76.2
Overall performance (%)	81.03		

THREE-CLASS COMPARISON. The classifier results show that the PD group is clearly separable from the other two disease groups. On the other hand, the PSP and MSA groups seem to overlap more strongly. We observe that the majority of the misclassification for both the PSP and MSA belong to either classes, which shows that these two groups are quite similar. In fact, it is known that PSP and MSA are hard to distinguish because the patients with either disorders show similar reduction in striatal and brain stem volumes [Eckert et al., 2004].

VISUALIZATION OF THE DATA POINTS. The scatter plots show the training data points with respect to their projections on the two leading eigenvectors of the relevance matrix. It can be observed in Fig. 4.17(a) that the PSP and healthy groups are clearly separable from the rest of the groups. But a small overlap exists between the PD and MSA groups even in the training set. Meanwhile, the three-class comparison in Fig. 4.17(b) shows a clear separation among the disease groups. This is encouraging since we are generally interested in distinguishing among the parkinsonian syndromes.



(a) Four class problem; 1=HC, 2=PD, 3=PSP, 4=MSA



(b) Three class problem; 1=PD, 2=PSP, 3=MSA

Figure 4.17: The visualization of the training data with respect to their projections on the two leading eigenvectors of the relevance matrix as observed in a single run of GMLVQ training.

4.3.2 *Support Vector Machine (SVM)*

Next we show the results of the leave-one-out cross validation of the SVM classifier for the different groups, both in a binary and multi-class comparison. Note that, as before, a subject is left out before the SSM/PCA process.

4.3.2.1 *Binary Classification*

Here, the classifier was used to separate each disease group from the healthy control group to determine its classification performance. As seen in Table 4.4, apart from the PD vs HC

Table 4.4: SVM classifier LOOCV performance for the different data sets (patients vs healthy controls, number of cases in brackets). The column Perf.(%) indicates the percentage of subject cases correctly classified per group, Sensitivity (%) the percentage of correctly classified patients, and Specificity (%) the percentage of correctly classified healthy controls.

Feature set(size)	Perf. (%)	Sensitivity (%)	Specificity (%)	AUC
PD-HC (38)	76.3	75	77.8	0.84
MSA-HC (39)	94.9	90.5	100	0.97
PSP-HC (35)	91.4	88.2	94.4	0.92

comparison, the other groups' performances improve in comparison to GMLVQ (cf. Table 4.1). However, the AUC measures for MSA and PSP are lower than those of GMLVQ, indicating that it outperforms the SVM when choosing an appropriate class bias to modify the nearest prototype classification. In comparison to the linear SVM in [Mudali et al., 2015], the results differ because different features have been used. Furthermore, here the LOOCV is done correctly by removing the test subject from the training set before applying the SSM/PCA method, whereas in [Mudali et al., 2015] the SSM/PCA method was applied to all subjects to obtain the scores before the LOOCV was performed.

4.3.2.2 *Multi-class Classification*

We also applied SVM to the multi-class datasets to determine its performance on larger datasets.

FOUR-CLASS COMPARISON. This involved the comparison of all the four groups, i.e., HC, PD, PSP, and MSA. In Table 4.5, the SVM four-group classification accuracy is slightly above chance level and lower than that of GMLVQ (see Table 4.2). But the classifier can separate the MSA group from the rest of the groups with an accuracy of 81%.

Table 4.5: Four-class problem: The confusion matrix and the overall performance of the SVM in the LOOCV scheme.

SVM classification	HC	PD	PSP	MSA
HC(18)	**12**	3	2	0
PD(20)	4	**12**	1	3
PSP(17)	1	2	**9**	5
MSA(21)	0	2	2	**17**
Class accuracy (%)	66.7	60	52.9	81.0
Overall performance (%)	65.8			

THREE DISEASE GROUPS. This involved the comparison of only the disease groups, i.e., PD, PSP and MSA without the healthy group. The separation of the disease groups using

Table 4.6: Three-class problem: The table shows the confusion matrix with the number of subject images correctly classified by the SVM for each class in bold and the overall LOOCV performance in percentage.

SVM classification	PD	PSP	MSA
PD(20))	**17**	1	2
PSP(17)	2	**10**	5
MSA(21)	3	2	**16**
Class accuracy (%)	85	58.8	76.2
Overall performance (%)	74.1		

SVM yields a better performance accuracy than the separation of the four groups (including the healthy group). Also, as in the GMLVQ classification, the PD group appears to be well separated from PSP and MSA.

4.4 DISCUSSION AND CONCLUSION

Both GMLVQ and SVM were studied and tested for the binary and multi-class problems. In the binary classification, GMLVQ performs better than SVM in the PD vs HC comparison (performance of 81.6%), but both achieve the same sensitivity of

75%. However, SVM performs better in the MSA vs HC and PSP vs HC comparisons. For the two-class problems we also considered the area under the curve (AUC) of the ROC, as it does not depend on the choice of a particular working point (threshold, class bias) in the classifier. In terms of the AUC, GMLVQ was seen to outperform or equal the performance of the SVM classifier. Additionally, in the multi-class problems, GMLVQ achieves a better accuracy than SVM.

The GMLVQ relevance matrix, which makes use of an adaptive weighting of features according to their discriminative power, displayed overall superior classification performance. In particular, for the PD vs HC comparison which has been challenging to discriminate using decision trees, GMLVQ was able to separate PD from HC with an accuracy of 81.6%, better than SVM by a margin of 5.3%. Although SVM classification performance for the MSA vs HC and PSP vs HC comparisons is better than GMLVQ, the AUC measures show that GMLVQ achieves superior binary classification of the distinct groups. Overall, GMLVQ also achieves a better accuracy for the multi-class classification. In addition, when it comes to explaining the results to the physicians, GMLVQ is more intuitive than SVM. The analysis of the resulting relevance matrix allows for the identification of particularly relevant features and combinations of features. LVQ methods parameterize the classifier in terms of prototypes, i.e. in terms of objects which are defined in the feature space. They can be interpreted as typical representatives of the classes and facilitate discussions with the domain experts. The relevance matrix in GMLVQ provides further insights into the structure of the classification problem as its elements quantify the importance of single features and pairs of features. This is in contrast to many other classification schemes, e.g. the Support Vector Machine, which do not offer the same degree of direct interpretability in feature space. These results should trigger further investigations from the clinical perspective.

It is interesting to compare our method with that of [Rauber et al., 2015], in which relevant features are selected to construct effective classifiers. By contrast, in the GMLVQ approach the features are not pre-selected or reduced but are weighted according to their discriminative power via the relevance matrix. The contribution of individual features to the classification process varies in accordance to their weight.

Clearly, the number of cases in the available data set is fairly small and our findings could be partly skewed by the small sample size. For instance, leave-one-out validation schemes are known to frequently yield unreliable estimates of performance. It is also possible that the performance of decision trees in [Mudali et al., 2015], which was found inferior to GMLVQ and SVM, might improve significantly for larger data sets (see comparable work in [Westenberg and Roerdink, 2002]). We intend to extend our work in this direction as more data become available in the future. Moreover, variants of the considered classifiers could be considered, e.g., SVM with more powerful kernels or LVQ systems with several prototypes per class or local distance matrices [Schneider et al., 2009].

5

DIFFERENTIATING EARLY AND LATE STAGE PARKINSON'S DISEASE PATIENTS FROM HEALTHY CONTROLS

ABSTRACT: *Parkinson's disease (PD) is a neurodegenerative disease which is difficult to diagnose at early disease stages. Brain imaging techniques like [18F]-fluorodeoxyglucose positron emission tomography (FDG-PET) may aid to identify disease-related changes in cerebral glucose metabolism. The scaled subprofile model with principal component analysis (SSM/PCA) is applied to FDG-PET data to extract features and corresponding patterns of glucose metabolism which can be used to distinguish PD subjects from healthy controls. From a previous study, the decision tree (DT) classifier's performance to separate the PD group from healthy controls was below chance level. This could be attributed to the small number of subjects in the dataset, combined with the early disease progression. In this study, we make use of an additional PD dataset, consisting of subject brain images obtained at a later disease stage. The features extracted by the SSM/PCA method are used for distinguishing PD subjects from healthy controls using three classification methods, that is, decision trees, Generalized Matrix Learning Vector Quantization (GMLVQ), and Support Vector Machine (SVM) with linear kernel. The classifiers are validated to determine their capability of classification given new subject data. We compare the classifiers' performances on the distinct early-stage and late-stage datasets, as well on the combined datasets. We also use the early and late-stage datasets interchangeably for training and testing the classifiers. We find that the DT classification performance on the late-stage dataset is considerably better than in the previous study, where we used early-stage data. For early-stage patients, the application of the GMLVQ and SVM classifiers gives a significant improvement as compared to the DT classifier.*

KEYWORDS: *Parkinson's disease, SSM/PCA, decision tree classification.*

5.1 INTRODUCTION

Parkinson's disease (PD) and other parkinsonian disorders such as progressive supranuclear palsy (PSP) and multiple system atrophy (MSA) often show overlap in symptoms at an early disease stage. An accurate diagnosis can only be achieved after long-term serial assessment by a movement disorders specialist [Hughes et al., 2002; Osaki et al., 2004]. This is problematic because early diagnosis is important for selecting appropriate treatments. We use typical patterns of glucose metabolism delineated by [18F]-Fluoro-deoxyglucose (FDG) PET with the purpose of differentiating between parkinsonian syndromes. Such patterns are extracted by applying the scaled subprofile

model and principal component analysis (SSM/PCA, [Moeller et al., 1987]) to FDG PET data of healthy controls and patients [Eidelberg, 2009]. The expression of previously identified patterns can be computed from the scans of new individuals. These pattern expression values are useful markers for disease [Niethammer and Eidelberg, 2012].

The decision tree method [Quinlan, 1993] was used in the previous study [Mudali et al., 2015] to classify parkinsonian syndromes. However, it was quite a challenge to separate the PD subjects from the healthy controls. This could be because the number of subjects in the dataset was not sufficient enough to train a robust decision tree classifier.

In this study, in addition to the dataset of early stage PD and healthy controls used in [Mudali et al., 2015], a larger dataset consisting of brain images of healthy controls and patients with PD obtained at a later disease stage is also used. It is desirable to generate a large dataset consisting of brain data obtained at all stages of disease progression to extract features which can be used to train a robust classifier. Therefore, we will investigate whether features that are more suitable to separate the PD and healthy groups can be extracted from the advanced disease stage dataset, showing evident disease patterns in the data; in other words, to extract patterns which are evidently associated with PD [Eckert et al., 2007].

In our earlier study [Mudali et al., 2015] the number of subjects was too small to separate the dataset in a training and test set to assess classifier accuracy. Therefore to estimate classification performance the Leave-One-Out Cross Validation (LOOCV) method was used. However, as is well known, the LOOCV performance results are only an indication of what can be achieved when training and test sets are defined by different input data. Since we now have independent PD data from different sources we can use one as training and the other as test set to determine classifier accuracy. For comparison with earlier results we also compute LOOCV performance for the case of single datasets.

The scaled subprofile model with principal component analysis (SSM/PCA) method [Moeller et al., 1987; Moeller and Strother, 1991] is used to extract the discriminative features from the brain data. Based on these features, the C4.5 decision tree classification algorithm [Quinlan, 1993] is used for building classifiers to separate the PD group from healthy controls.

Decision trees have the advantage of being easily constructed and understood, hence they provide an insight into the most important features for classification [Al Snousy et al., 2011].

In previous brain imaging studies several other classifiers have been used with promising results. An example is the Support Vector Machine (SVM) which has been used to detect various neurological and psychiatric diseases [Magnin et al., 2009; Haller et al., 2012; Orrù et al., 2012]. Another example is Generalized Matrix Learning Vector Quantization (GMLVQ), which has been used in many disciplines including image analysis and bioinformatics, see Yeo et al. [2015]. A strong point of the prototype-based GMLVQ classifier is that it is intuitive and easy to interpret. In addition, it provides insight into the relevance of individual features for the classification [Schneider et al., 2009]. For this reason, in addition to the decision tree method, we applied the SVM and GMLVQ classifiers to the subject scores extracted from the FDG-PET brain image data, with the aim to study classification accuracy of the methods given larger and different datasets.

5.2 METHOD

5.2.1 *Subjects*

Subject brain images were acquired from two hospitals. First, data of forty nine patients diagnosed with PD according to the UK Parkinson's Disease Society Brain Bank criteria were obtained from the Movement Disorders Unit of the Clinica Universidad de Navarra (CUN), Spain. Clinical and metabolic data of these patients was previously published in García-García et al. [2012]. In addition, 19 age-and gender-matched control subjects without a history of neurologic, psychiatric illness and no abnormalities on MRI were included. From the 49 PD subjects, we randomly selected 20 PD subjects for training the classifier (PD subjects of dataset D1_CUN, see table 5.2), and 29 for testing the classifier (PD subjects of dataset D2_CUN/UMCG see table 5.2). Age, gender, disease duration, Unified Parkinson's Disease Rating Scale (UPDRS) motor ratings and Hoehn & Yahr (H&Y) scores did not differ significantly between PD patients in the two cohorts. Ethical permission for the procedures was obtained from the Ethics Committee for Medical Research of the University of Navarra. Written consent was obtained at

each institution from all subjects following detailed explanation of the testing procedures. All the 19 healthy controls were added to the training set to make a total of 39 subjects (dataset D1_CUN).

Second, 20 PD subjects and 18 healthy controls were obtained from the University Medical Center Groningen (UMCG), more details are found in [Teune et al., 2010]. The 18 healthy controls (from UMCG) were added to the test set of 29 PD (dataset D2_CUN/UMCG, see table 5.2) from CUN to make 47 subjects. These 18 HC subjects from UMCG were considered for the test set because the 19 HC from CUN were too few to divide into the training and test sets. Also, the 20 PD and the earlier mentioned 18 healthy controls both from Teune et al. [2010] (dataset D3_UMCG, see table 5.2) were considered for training and testing the classifiers. This particular dataset D3_UMCG was obtained at an early disease stage.

The original datasets from the University Medical Center Groningen (UMCG) and the Clinica Universidad de Navarra (CUN) are shown in Table 5.1;

Table 5.1: The original datasets as provided from their respective sources.

Subjects	**Source**
49 PD and 19 HC	CUN
20 PD and 18 HC	UMCG [Teune et al., 2010]

The following table 5.2 shows the arrangement of the derived datasets from the original datasets for experiments, i.e., for training and testing classifiers.

Table 5.2: The arrangement of the datasets as used for both training and testing of classifiers.

Dataset	**Description**
D1_CUN	20 PD & 19 HC both groups from CUN
D2_CUN/UMCG	29 PD from CUN & 18 HC from UMCG
D3 _UMCG	20 PD & 18 HC both groups from UMCG

5.2.2 *Image acquisition and preprocessing*

The CUN subjects were scanned with [18F]fluorodeoxyglucose Positron Emission Tomography (FDG-PET) under resting conditions. Patients were studied in the 'on' pharmacological condition (under the effect of anti-parkinsonian medication). Central nervous system depressant drugs were withdrawn, and subjects fasted overnight before FDG-PET scanning. FDG-PET imaging was performed in 3D mode using a Siemens ECAT EXAT HR+ scanner (Siemens, Knoxville, TN). Image acquisition was performed in a resting state with the subject's eyes closed in a dimly lighted room with minimal auditory stimulation. Images were reconstructed by means of a filtered back-projection method using ECAT software (version 7.2; Siemens). Preprocessing of imaging data was performed by SPM8 software (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK) implemented in Matlab 8.0 (Mathworks Inc, Sherborn, MA). All images were spatially normalized onto a PET template in Montreal Neurological Institute (MNI) brain space and then smoothed by a Gaussian filter of 10 mm FWHM. The UMCG FDG-PET brain data was scanned as described previously by Teune et al. [2010] and preprocessed in the same way as the CUN data.

5.2.3 *Feature extraction, classification and classifier validation*

The same steps as those of Mudali et al. [2015] were followed to extract features from the brain image data in the form of subject scores on principal components using the SSM/PCA method [Moeller et al., 1987; Moeller and Strother, 1991]. These subject scores were the features provided to the decision tree inducer, GMLVQ and SVM to train and test the classifiers for the different cohorts. All the extracted features were considered for building the classifiers. The classifiers' performance was determined using leave-one-out cross validation (LOOCV). In each LOOCV iteration, a subject was removed from the training set before the SSM/PCA process to obtain features for training the classifiers. The left-out-subject was then used for testing the trained classifier.

In anticipation of better classification performance, we used the dataset D1_CUN which was obtained at a later disease stage to train the classifier. Then we tested the classifier using

the subject scores extracted from both dataset D2_CUN/UMCG (PD group obtained at a later disease stage) and D3_UMCG (the PD subjects obtained at an earlier disease stage and healthy controls), see Table 5.2.

The decision tree classifiers are built using the C4.5 decision tree algorithm designed by Quinlan [1993]. This algorithm takes subject scores as inputs and outputs corresponding decision trees as classifiers [Mudali et al., 2015]. Additionally, we use the gmlvq-toolbox by Biehl [2015] to train and test GMLVQ classifiers with default parameters. Further, for the SVM we use the linear kernel since the dataset is still small. The Matlab R2014a functions "fitcsvm" and "predict" are used to train and test the classifiers respectively, see Mudali et al. [2016b].

5.3 CLASSIFICATION RESULTS

The data was used to train and test three different types of classifiers i.e., decision trees (DT), GMLVQ and SVM. The LOOCV results and the performances of training the classifiers on one cohort and testing on another are included.

5.3.1 *Classifier Leave-one-out cross validation (LOOCV) on dataset D1_CUN*

In this section we present the results obtained after the LOOCV of the DT, GMLVQ and SVM classifiers on dataset D1_CUN (39 subjects).

Table 5.3: GMLVQ, SVM, and DT LOOCV performance: Perf. = total accuracy, Sens. = Sensitivity and Spec. = Specificity with respect to detecting the disease.

Classifiers	GMLVQ	SVM	DT
Sens.%	100	100	90
Spec.%	89.5	94.7	84.2
Perf.%	94.9	97.4	87.2

Although both GMLVQ and SVM outperform DT in the LOOCV of dataset D1_CUN (the training set), the DT classifier is competitive since the difference in the performances is relatively small as can be seen in Table 5.3. We observe that with the

CUN dataset of PD subjects obtained at a later disease stage, the DT is capable of separating the groups to a satisfactory extent.

5.3.2 *GMLVQ, SVM and DT performance with dataset D1_CUN as the training set and D2_CUN/UMCG as the test set*

Here we used the later disease stage dataset D1_CUN for training and dataset D2_CUN/UMCG for testing, which contains advanced PD subjects from CUN and a HC group from UMCG.

Table 5.4: GMLVQ, SVM, and DT performance. D1_CUN as the training set and D2_CUN/UMCG as the test set: The table shows the confusion matrix for the classification of dataset D2_CUN/UMCG with the overall performance (perf.) in percentage.

	GMLVQ		**SVM**		**DT**	
	HC	PD	HC	PD	HC	PD
HC (18 subjects)	**14**	4	**14**	4	**12**	6
PD (29 subjects)	3	**26**	2	**27**	3	**26**
Class accuracy (%)	77.8	89.7	77.8	93.1	66.7	89.7
Overall perf. (%)	85.1		87.2		80.9	

As can be seen in table 5.4, for DT only 3 out of 29 PD subjects from dataset D2_CUN/UMCG are misclassified as healthy controls with an overall performance of 80.9%. With respect to GMLVQ and SVM, the only difference is in the PD group, where SVM correctly classifies just one more PD subject than GMLVQ. However, both GMLVQ and SVM perform better than DT due to a higher accuracy on the HC group.

5.3.3 *Classifier performance with dataset D1_CUN as the training set and D3_UMCG as the test set*

In the setting discussed in this subsection, the training and test sets are from two different sites, that is, dataset D1_CUN is used for training and D3_UMCG for testing. The classifier results are shown in Table 5.5;

The DT performance as seen in Table 5.5 is lower than that of 80.9% in Table 5.4 when testing with dataset D2_CUN/UMCG. However, it is higher than the PD group performance of 63.2%

Table 5.5: GMLVQ, SVM, and DT performance. D1_CUN as the training set and D3_UMCG as the test set: The table shows the confusion matrix for the classification of dataset D3_UMCG with the overall performance in percentage.

	GMLVQ		**SVM**		**DT**	
	HC	PD	HC	PD	HC	PD
HC (18 subjects)	**14**	4	**14**	4	**12**	6
PD (20 subjects)	6	**14**	6	**14**	7	**13**
Class accuracy (%)	77.8	70	77.8	70	66.7	65
Overall performance (%)	73.7		73.7		65.8	

in Mudali et al. [2015]. Again this means that the decision tree classifier's ability to separate the two groups has improved. On the other hand, both GMLVQ and SVM register the same performance of 73.7% which is better than that of DT.

5.3.4 *Classifier performance with dataset D3_UMCG as the training set and D1_CUN as the test set*

The setting in this subsection is the reverse of that in subsection 5.3.3. At first sight, it may seem surprising to use the early-stage data set for training. Our motivation for this experiment is to see whether the early-stage data perhaps already do contain some features that help to differentiate PD subjects from healthy controls.

Table 5.6: GMLVQ, SVM, and DT performance. Dataset D3_UMCG as the training set and D1_CUN as the test set: The confusion matrix and the overall performance in percentage.

	GMLVQ		**SVM**		**DT**	
	HC	PD	HC	PD	HC	PD
HC (19 subjects)	**18**	1	**18**	1	**16**	3
PD (20 subjects)	2	**18**	2	**18**	10	**10**
Class accuracy (%)	94.7	90	94.7	90	84.2	50
Overall performance (%)	92.3		92.3		66.7	

Using D3_UMCG to train the classifiers and testing with D1_CUN yielded the same performance of 92.3% for both GMLVQ and SVM which is better than 66.7% for DT. It is interesting that the performance is better than in the setting of

section 5.3.3 (training with the CUN dataset and testing with respect to the UMCG dataset).

5.3.5 *LOOCV of the combined datasets D1_CUN and D3_UMCG*

Datasets D1_CUN and D3_UMCG were combined to make a dataset of 77 subjects arranged into *two classes*, i.e., 37 HC and 40 PD subjects. The GMLVQ classifier performance on the combined dataset was validated using the leave-one-out method, so as to determine the capability of the classifier to distinguish between the PD and HC subjects. Table 5.7 shows the confusion matrix for the two-class problem.

Table 5.7: GMLVQ, SVM, and DT LOOCV performance of the combined datasets D1_CUN and D3_UMCG in two classes: The confusion matrix and the overall performance.

	GMLVQ		**SVM**		**DT**	
	HC	PD	HC	PD	HC	PD
HC (37 subjects)	**35**	2	**32**	5	**24**	13
PD (40 subjects)	4	**36**	4	**36**	10	**30**
Class accuracy (%)	94.6	90	86.5	90	64.9	75
Overall performance (%)	92.2		88.3		70.1	

The GMLVQ classifier can separate the two groups with a 92.2% accuracy as seen in Table 5.7, with sensitivity of 90% and specificity of 94.6%. SVM is fairly competitive, with a clearly lower performance of DT.

Having obtained good GMLVQ accuracy in Table 5.7, we next applied the GMLVQ classifier where we arranged the data from the D1_CUN and D3_UMCG datasets into *four distinct classes*, i.e., 18 HC from UMCG, 20 PD from UMCG, 19 HC from CUN, and 20 PD from CUN. This was done in anticipation of the GMLVQ classification accuracy in separating the CUN subjects from the UMCG subjects. The results for the four-class problem are shown in Table 5.8.

In Table 5.8, the GMLVQ classifier is able to separate all the CUN PD subjects from the rest of the subjects. However, 8 out of 20 UMCG PD subjects are misclassified as UMCG HC (5 subjects) and CUN PD (3 subjects).

Table 5.8: GMLVQ LOOCV performance on the combined datasets D1_CUN and D3_UMCG in four classes: The table shows the number of test subject images correctly classified for each class (in bold) with the overall performance in percentage.

	CUN HC	UMCG HC	CUN PD	UMCG PD
CUN HC (19 subjects)	**17**	1	1	0
UMCG HC (18 subjects)	1	**17**	0	0
CUN PD (20 subjects)	0	0	**20**	0
UMCG PD (20 subjects)	0	5	3	**12**
Class accuracy (%)	89.5	94.4	100	60
Overall performance (%)	85.7			

5.4 DISCUSSION AND CONCLUSION

This study has focused on the differentiation between Parkinson's disease and healthy control brain patterns. In the previous study by Mudali et al. [2015], the decision tree (DT) classifier displayed relatively poor classification performance as assessed by leave-one-out cross validation (LOOCV). This poor performance was attributed to the small number of subjects in the dataset used and/or the brain data being obtained at an early disease stage.

The present study shows that one can obtain high LOOCV performances for patients at a more advanced disease stage using different classifiers; see Table 5.3 for the D1_CUN data. Although GMLVQ and SVM reach the highest performance, the decision tree classifier also performs very well. It reaches a performance around 87%, which is a significant improvement with respect to the results in Mudali et al. [2015], which were obtained for the D3_UMCG data. The difference between these data sets is not the number of subjects, but the fact that the D1_CUN data set corresponds to a later disease stage, with more metabolic changes than the early disease stage dataset. Hence, the disease pattern is more pronounced and the extracted features apparently are more informative with respect to separating the late-stage PD subjects from healthy controls.

The availability of a data set from CUN with a larger number of subjects, as well as data sets from different sites (i.e., CUN and UMCG), allowed us to perform a number of additional tests. When D1_CUN was used as the training set and D2_CUN/UMCG as the test set (in both data sets the PD sub-

jects are from CUN), the performances of GMLVQ and SVM was still very good (85% and 87%, resp), while with 81% the DT was still competitive; see Table 5.4.

When D1_CUN was used as the training set and D3_UMCG as the test set (now the PD subjects are from CUN and UMCG, respectively), the performances are significantly lower for all classifiers; see Table 5.5. Comparing the results with Table 5.4 we see that the main reason is the higher percentage of PD subjects in the test set that are misclassified as healthy controls. As before, the explanation is that in this experiment the PD subjects in the test set are early stage patients from UMCG, which are hard to distinguish from healthy controls.

Somewhat surprisingly, training with the early-stage UMCG data and testing with respect to the late-stage CUN data yields much better performance than *vice versa* for the GMLVQ and SVM classifiers, as can be observed by comparing Tables 5.5 and 5.6. Training on early stage data with the GMLVQ and SVM classifiers seems to infer the subtle differences between early-stage PD and HC subjects, which then can be successfully used for the distinguishing late-stage PD from HC. Although in late stage data the differences between PD and HC will be more pronounced, training on such data is apparently less effective for classification when the test set contains early-stage patients which are quite similar to healthy controls. For the DT classifier, no significant improvement is seen when comparing Tables 5.5 and 5.6. The decision tree needs to take separate decisions on many features, while GMLVQ and SVM can handle linear combinations of feature values. For the DT this leads to the problem of overfitting and limited generalizability, especially when the number of subjects is relatively small. With pruning (feature selection) the overfitting problem can be reduced, but at the cost of lower performance; see [Mudali et al., 2015] for a more extensive analysis.

The fact that early-stage PD subjects from the UMCG dataset are closer to healthy controls than to the late-stage PD samples in the CUN dataset can also clearly be inferred from Table 5.8. In this 4-class problem, the CUN PD subjects can be perfectly identified (no misclassification errors). Most errors occur for UMCG PD subjects that are misclassified as UMCG HC. Also, three of the PD UMCG subjects are misclassified as CUN PD subjects, suggesting that these three subjects are closer to late-stage than early-stage patients. Therefore it could be interesting

to explore in detail the relationship between these three PD UMCG subjects and the CUN subjects and, in particular, to extensively study the corresponding subject brain images.

On the combined datasets D1_CUN and D3_UMCG (2-class problem, all PD combined, all HC combined), the classifiers are also able to differentiate between Parkinson's disease and healthy controls with good performances as seen in Table 5.7, especially for the GMLVQ and SVM classifiers.

In conclusion, this study has shown that by applying state-of-the-art classifiers to FDG-PET brain data, Parkinson's disease subjects can be separated from the healthy controls with high accuracy. We have shown that the application of the GMLVQ and SVM classifiers can give a significant improvement as compared to the DT classifier, especially for classifying early-stage patients.

With respect to understanding the behaviour of the classification methods, the GMLVQ and DT methods have proven to be more intuitive than SVM. Moreover, they can handle computationally very large feature sets. When both high accuracy and intuitive understanding of the classifier is desired, the GMLVQ method can be recommended.

We expect that the classifier performance will show further improvement, even for early-stage brain data, when the number of subjects in the training dataset further increases. This is the ultimate goal of the GLIMPS project [Teune et al., 2012], which aims at establishing a large database of FDG-PET scans from the Netherlands and abroad.

6 SUMMARY AND CONCLUSIONS

6.1 SUMMARY AND DISCUSSION

Neurodegenerative diseases continue to be a challenge in the developed society where the life expectancy is high. If measures are not put in place, these diseases will continue to affect the elderly and increase the mortality rate. Since they progress slowly, they are not easy to diagnose at an early stage. Moreover, they portray similar disease features, which makes them hard to differentiate.

In this thesis, our objective was to devise techniques to extract biomarkers from FDG-PET brain data for the prediction and classification of neurodegenerative diseases, in particular parkinsonian syndromes. Therefore we used Principal Component Analysis (PCA) in combination with the scaled subprofile model (SSM) to extract features from the brain data to classify these disorders. Furthermore we validated the classifiers.

A background to neurodegenerative diseases and brain imaging techniques was given in chapter 1. In chapter 2 we started our study of classification of parkinsonian syndromes using decision trees because they are easy to understand. Features in the form of subject scores were extracted from the FDG-PET brain data using the SSM/PCA method. These features were input to the C4.5 decision tree inducer to train classifiers. The classifiers were validated with a leave-one-out method. The results showed that the decision tree can separate the MSA and PSP subjects from healthy controls but is unsuccessful when it comes to the PD subjects, although DT accuracy improved after reducing the number of features to include only the most relevant ones in the classification process. Therefore, we concluded that since the FDG-PET activation pattern of PD patients (especially at an early disease stage) and healthy controls is similar, the two groups are hard to separate. Additionally, the size of the data sets used in this chapter was too small to achieve better results. Pairwise comparisons of disease groups (without the healthy group) yielded a better

classification performance. Other classifiers like LDA, nearest neighbors, CART, random forests, etc., were also applied to the same data but they were not exceptional in terms of the classification. The decision trees also helped us to visualise the classification results, hence providing an insight into the distribution of features.

In chapter 3, we compared the decision tree method to the stepwise regression (SR) method which aims at linearly combining a few "good" PCA components. The SR method performed better than the DT method in the classification of the parkinsonian syndromes. This is because the SR method combines the best features into one robust feature for classifying the Parkinsonian syndromes, unlike the DT method which uses the features individually. Interestingly, we found that when the same robust feature is provided to the DT inducer as input, the accuracy is equally high. Therefore, combining the two methods, i.e., combining features using the SR procedure and providing them to the DT method for classifying the syndromes is feasible. An advantage of the DT classifier is that it can be applied to multi-class problems, unlike the stepwise regression method.

The decision tree method was our initial option for the classification of neurodegenerative diseases due to the fact that decision trees are intuitive and easy to understand. However, having obtained not completely satisfactory results in the previous chapters, we opted to try other classification methods in chapter 4. In this chapter, we applied the GMLVQ and SVM classifiers to Parkinsonian syndrome data in the hope to achieve better classification results. As before, we used the SSM/PCA method to obtain the features for classification and supplied them to the GMLVQ and the SVM classifiers to classify the subject image data. The results show that both GMLVQ and SVM are better than the DT method in the classification of early-stage parkinsonian syndromes. SVM fairly competes with GMLVQ in the binary classification, but the Area Under the Curve (AUC) measures show that GMLVQ is superior. With the multi-class problems, GMLVQ achieved a better classification accuracy than SVM. Unlike SVM, with GMLVQ the results are easier to interpret in the form of the diagonal and off-diagonal matrices.

After acquiring a dataset with a larger number of PD scans, moreover at a later disease stage, we applied the decision tree, GMLVQ and SVM classification methods to this data. In chap-

ter 5, we compared the different classifier performances regarding the separation of the 'late-stage' PD group from the healthy controls. The decision tree leave-one-out cross validation performance results for this particular group (advanced stage) are far better than those in chapter 2. Furthermore, the GMLVQ and SVM perform much better than the decision tree in the separation of the PD and HC groups for early disease-stage patients. On the other hand, GMLVQ and DT are more intuitive than SVM. They both can handle very large feature sets. We found out that training and testing using a bigger dataset including late-stage PD brain images yields much better results than the smaller size dataset with early-stage PD scans only. Therefore, large training datasets aid in better classification of neurodegenerative diseases. All the classification methods used in this thesis performed well with the later disease stage data. We conclude that GMLVQ and decision tree methods can be recommended for further research on neurodegenerative disease classification and prediction.

6.2 FUTURE WORK

In chapter 2 the decision tree method was applied to small size datasets. It is important to generate more data and apply the decision tree method to larger size datasets in anticipation of better results. Moreover, data can encompass different imaging modalities like MRI, fMRI, DTI, etc. Also, it is important to visualise the decision tree diagrams to look critically at the thresholds used in the classification. By doing so, the disease stage can be determined. This can be achieved by displaying the subject scores (on the features chosen by the classifier) on scatter plots or histograms to determine the distance of subject scores from the thresholds. Additionally, it could be interesting to explore the decision trees using interactive visualisation techniques.

Since in chapter 3 we see an outstanding performance of the stepwise regression (SR) procedure in the separation of two groups, a deeper look into the extension of the SR procedure to include the comparison of more than two groups is interesting.

In chapter 5, the leave-one-out cross validation of the combined early/late disease stage data yielded better results. Therefore, we propose that large datasets of subjects at several phases of disease progression should be accumulated to aid in building

robust classifiers. These classifiers can then be used to predict the types and stages of neurodegenerative diseases.

Finally, in addition to subject scores, other types of features could be generated from FDG-PET or MRI data to improve the classification of parkinsonian syndromes.

BIBLIOGRAPHY

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

M. S. Al-Rawi and J. P. S. Cunha. Using permutation tests to study how the dimensionality, the number of classes, and the number of samples affect classification analysis. In *Image Analysis and Recognition*, pages 34–42. Springer, 2012.

M. B. Al Snousy, H. M. El-Deeb, K. Badran, and I. A. Al Khlil. Suite of decision tree-based classification algorithms on cancer gene expression data. *Egyptian Informatics Journal*, 12(2): 73–82, 2011.

L. Appel, M. Jonasson, T. Danfors, D. Nyholm, H. Askmark, M. Lubberink, and J. Sörensen. Use of 11C-PE2I PET in differential diagnosis of parkinsonian disorders. *Journal of Nuclear Medicine*, 56(2):234–242, 2015. doi> 10.2967/JNUMED.114.148619

J. Ashburner and K. Friston. Voxel-based morphometry - the methods. *Neuroimage*, 11(6):805–821, 2000.

P. J. Basser, J. Mattiello, and D. LeBihan. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1):259–267, 1994.

D. Berg. Biomarkers for the early detection of Parkinson's and Alzheimer's disease. *Neurodegenerative Diseases*, 5(3-4): 133–136, 2008.

M. Biehl. A no-nonsense Matlab (TM) toolbox for GMLVQ, 2015. Software available at `http://www.cs.rug.nl/biehl/gmlvq.html`.

M. Biehl, B. Hammer, F.-M. Schleif, P. Schneider, and T. Villman. Stationarity of matrix relevance LVQ. In *Proc. International Joint Conference on Neural Networks, IJCNN 2015, Killarney / Ireleand*. IEEE, 2015. doi> 10.1109/IJCNN.2015.7280441

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl. Limited rank matrix learning, discriminative dimension reduction and visualization. *Neural Networks*, 26: 159–173, 2012.

C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

E. J. Burton, I. G. McKeith, D. J. Burn, E. D. Williams, and J. T. O'Brien. Cerebral atrophy in Parkinson's disease with and without dementia: a comparison with Alzheimer's disease, dementia with Lewy bodies and controls. *Brain*, 127(4):791–800, 2004.

R. P. Carne, T. J. O'Brien, C. J. Kilpatrick, L. R. MacGregor, L. Litewka, R. J. Hicks, and M. J. Cook. 'MRI-negative PET-positive' temporal lobe epilepsy (TLE) and mesial TLE differ with quantitative MRI and PET: a case control study. *BMC neurology*, 7(1):16, 2007.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

G. Chételat, B. Desgranges, B. Landeau, F. Mézenge, J. Poline, V. De La Sayette, F. Viader, F. Eustache, and J.-C. Baron. Direct voxel-based comparison between grey matter hypometabolism and atrophy in Alzheimer's disease. *Brain*, 131(1):60–71, 2008.

T. W. Chow, D. C. Mamo, H. Uchida, A. Graff-Guerrero, S. Houle, G. S. Smith, B. G. Pollock, and B. H. Mulsant. Test-retest variability of high resolution positron emission tomography (PET) imaging of cortical serotonin (5HT2A) receptors in older, healthy adults. *BMC medical imaging*, 9(1): 12, 2009.

M. E. Cintra, M. C. Monard, and H. A. Camargo. FuzzyDT-A fuzzy decision tree algorithm based on C4.5. In *Proceedings of the Brazilian Congress on Fuzzy Systems*, pages 199–211, 2012.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi> 10.1007/BF00994018

N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

C. Davatzikos, S. M. Resnick, X. Wu, P. Parmpi, and C. M. Clark. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *Neuroimage*, 41(4): 1220–1227, 2008.

H. du Buf and M. M. Bayer, editors. *Automatic Diatom Identification*. World Scientific Publishing, Singapore, 2002.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000. ISBN 0471056693.

T. Eckert, M. Sailer, J. Kaufmann, C. Schrader, T. Peschel, N. Bodammer, H.-J. Heinze, and M. A. Schoenfeld. Differentiation of idiopathic Parkinson's disease, multiple system atrophy, progressive supranuclear palsy, and healthy controls using magnetization transfer imaging. *Neuroimage*, 21(1):229–235, 2004.

T. Eckert, A. Barnes, V. Dhawan, S. Frucht, M. F. Gordon, A. S. Feigin, and D. Eidelberg. FDG PET in the differential diagnosis of parkinsonian disorders. *Neuroimage*, 26:912–921, 2005.

T. Eckert, C. Tang, and D. Eidelberg. Assessment of the progression of Parkinson's disease: a metabolic network approach. *The Lancet Neurology*, 6(10):926–932, 2007.

T. Eckert, C. Tang, Y. Ma, N. Brown, T. Lin, S. Frucht, A. Feigin, and D. Eidelberg. Abnormal metabolic networks in atypical parkinsonism. *Movement Disorders*, 23(5):727–733, 2008. doi> 10.1002/MDS.21933

D. Eidelberg. Metabolic brain networks in neurodegenerative disorders: a functional imaging approach. *Trends in Neurosciences*, 32(10):548–557, 2009. doi> 10.1016/J.TINS.2009.06.003

M. T. Fodero-Tavoletti, R. Cappai, C. A. McLean, K. E. Pike, P. A. Adlard, T. Cowie, A. R. Connor, C. L. Masters, C. C. Rowe, and V. L. Villemagne. Amyloid imaging in Alzheimer's disease and other dementias. *Brain imaging and behavior*, 3(3): 246–261, 2009.

R. Frackowiak, K. Friston, C. Frith, R. Dolan, C. Price, S. Zeki, J. Ashburner, and W. Penny. *Human Brain Function*. Academic Press, 2nd edition, 2003.

K. J. Friston, J. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny, editors. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007.

K. Fukunaga. *Introduction to Statistical Pattern Recognition (2Nd Ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990. ISBN 0-12-269851-7.

M. Fukunda, J. Mentis, M, Y. Ma, V. Dhawan, A. Antonini, E. Lang, A, M. Lozano, A, J. Hammerstad, K. Lyons, C. Koller, W, R. Moller, J, and D. Eidelberg. Networks mediating the clinical effects of pallidal brain stimulation for Parkinson's disease. A PET study of resting-state glucose metabolism. *Brain*, 124:1601–1609, 2001.

D. García-García, P. Clavero, C. G. Salas, I. Lamet, J. Arbizu, R. Gonzalez-Redondo, J. A. Obeso, and M. C. Rodriguez-Oroz. Posterior parietooccipital hypometabolism may differentiate mild cognitive impairment from dementia in Parkinson's disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 39(11):1767–1777, 2012.

G. Garraux, C. Phillips, J. Schrouff, A. Kreisler, C. Lemaire, C. Degueldre, C. Delcour, R. Hustinx, A. Luxen, A. Destée, and E. Salmon. Multiclass classification of FDG PET scans for the distinction between Parkinson's disease and atypical parkinsonian syndromes. *NeuroImage: Clinical*, 2:883–893, 2013. doi> 10.1016/j.nicl.2013.06.004

S. Gilman, G. K. Wenning, P. A. Low, D. J. Brooks, C. J. Mathias, J. Q. Trojanowski, N. W. Wood, C. Colosimo, A. Dürr, C. J. Fowler, H. Kaufmann, T. Klockgether, A. Lees, W. Poewe, N. Quinn, T. Revesz, D. Robertson, P. Sandroni, K. Seppi, and M. Vidailhet. Second consensus statement on the diagnosis of multiple system atrophy. *Neurology*, 71(9):670–676, 2008. doi> 10.1212/01.wnl.0000324625.00404.15

W. Golder. Functional magnetic resonance imaging–basics and applications in oncology. *Onkologie*, 25(1):28–31, 2002.

P. Golland and B. Fischl. Permutation tests for classification: towards statistical significance in image based studies. In

Information Processing in Medical Imaging, volume 2732 of *Lecture Notes in Computer Science*, pages 330–341. Springer, 2003.

M. Greicius. Resting-state functional connectivity in neuropsychiatric disorders. *Current Opinion in Neurology*, 21(4):424–430, 2008.

A. E. Guttmacher, F. S. Collins, R. L. Nussbaum, and C. E. Ellis. Alzheimer's disease and Parkinson's disease. *New England Journal of Medicine*, 348(14):1356–1364, 2003.

S. Haller, S. Badoud, D. Nguyen, V. Garibotto, K. Lovblad, and P. Burkhard. Individual detection of patients with Parkinson disease using support vector machine analysis of diffusion tensor imaging data: initial results. *American Journal of Neuroradiology*, 33(11):2123–2128, 2012.

B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8–9):1059–1068, 2002.

B. Hammer, M. Strickert, and T. Villmann. Relevance LVQ versus SVM. In *Artificial Intelligence and Soft Computing-ICAISC 2004*, pages 592–597. Springer, 2004.

S. Hellwig, F. Amtage, A. Kreft, R. Buchert, O. H. Winz, W. Vach, T. S. Spehl, M. Rijntjes, B. Hellwig, C. Weiller, C. Winkler, W. A. Weber, O. Tüscher, and P. T. Meyer. [18F]FDG-PET is superior to [123I]IBZM-SPECT for the differential diagnosis of parkinsonism. *Neurology*, 79(13):1314–1322, 2012. doi> 10.1212/WNL.0B013E31826C1B0A

C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

A. J. Hughes, S. E. Daniel, Y. Ben-Shlomo, and A. J. Lees. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. *Brain*, 125(4):861–870, 2002.

M. Ito, H. Watanabe, Y. Kawai, N. Atsuta, F. Tanaka, S. Naganawa, H. Fukatsu, and G. Sobue. Usefulness of combined fractional anisotropy and apparent diffusion coefficient values for detection of involvement in multiple system atrophy. *J Neurol Neurosurg Psychiatry*, 78:722–728, 2006.

T. Johnsson. A procedure for stepwise regression analysis. *Statistical Papers*, 33(1):21–29, 1992.

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

I. Litvan, Y. Agid, D. Calne, G. Campbell, B. Dubois, R. C. Duvoisin, C. G. Goetz, L. I. Golbe, J. Grafman, J. H. Growdon, M. Hallett, J. Jankovic, N. P. Quinn, E. Tolosa, and D. S. Zee. Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome): Report of the NINDS-SPSP international workshop. *Neurology*, 47(1):1–9, 1996. doi> 10.1212/WNL.47.1.1

I. Litvan, K. P. Bhatia, D. J. Burn, C. G. Goetz, A. E. Lang, I. McKeith, N. Quinn, K. D. Sethi, C. Shults, and G. K. Wenning. SIC task force appraisal of clinical diagnostic criteria for parkinsonian disorders. *Movement Disorders*, 18(5):467–486, 2003. doi> 10.1002/MDS.10459

Y. Ma, T. Chenke, P. G. Spetsieris, V. Dhawan, and D. Eidelberg. Abnormal metabolic network activity in Parkinson's disease: test-retest reproducibility. *Journal of Cerebral Blood Flow & Metabolism*, 27(3):597–605, 2007.

Y. Ma, C. Tang, J. R. Moeller, and D. Eidelberg. Abnormal regional brain function in Parkinson's disease: truth or fiction? *NeuroImage*, 45(2):260–266, 2009. doi> 10.1016/J.NEUROIMAGE.2008.09.052

B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehéricy, and H. Benali. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology*, 51(2):73–83, 2009.

J. R. Moeller and S. C. Strother. A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *J Cereb Blood Flow Metab*, 11(2):A121–135, 1991.

J. R. Moeller, S. C. Strother, J. J. Sidtis, and D. A. Rottenberg. Scaled subprofile model: a statistical approach to the analysis of functional patterns in positron emission tomographic data. *J Cereb Blood Flow Metab*, 7(5):649–58, 1987.

J. R. Moeller, T. Ishikawa, V. Dhawan, P. Spetsieris, F. Mandel, G. E. Alexander, C. Grady, P. Pietrini, and D. Eidelberg. The metabolic topography of normal aging. *J Cereb Blood Flow Metab*, 16(3):385–98, 1996.

D. Mudali, L. K. Teune, R. J. Renken, K. L. Leenders, and J. B. T. M. Roerdink. Classification of Parkinsonian syndromes from FDG-PET brain data using decision trees with SSM/PCA features. *Computational and Mathematical Methods in Medicine*, Article ID 136921:1–10, 2015. doi> 10.1155/2015/136921

D. Mudali, M. Biehl, S. K. Meles, R. J. Renken, D. García-García, P. Clavero, J. Arbizu, J. Obeso, M. Rodriguez-Oroz, K. L. Leenders, and J. B. T. M. Roerdink. Differentiating early and late stage Parkinson's Disease patients from healthy controls using SSM/PCA features, 2016a. In preparation.

D. Mudali, M. Biehl, K. L. Leenders, and J. B. T. M. Roerdink. LVQ and SVM classification of FDG-PET brain data. In E. M. et al., editor, *Advances in Self-Organizing Maps and Learning Vector Quantization. Proc. WSOM 2016, 11th Workshop on Self-Organizing Maps*, number 428 in Advances in Intelligent Systems and Computing. Springer International Publishing Switzerland, 2016b. doi> 10.1007/978-3-319-28518-4_18

D. Mudali, L. K. Teune, R. J. Renken, K. L. Leenders, and J. B. T. M. Roerdink. Comparison of decision tree and stepwise regression methods in classification of FDG-PET brain data using SSM/PCA features. In *8th International Conference on Advanced Computational Intelligence, ICACI, Febr 14-16, Thailand*, 2016c.

A. P. Muniyandi, R. Rajeswari, and R. Rajaram. Network anomaly detection by cascading K-means clustering and C4.5 decision tree algorithm. In *Proceedings of the International Conference on Communication Technology and System Design 2011*, volume 30, pages 174–182. Procedia Engineering, 2012. doi> 10.1016/j.proeng.2012.01.849

M. Niethammer and D. Eidelberg. Metabolic brain networks in translational neurology: concepts and applications. *Annals of neurology*, 72(5):635–647, 2012.

G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, 36(4): 1140–1152, 2012.

Y. Osaki, Y. Ben-Shlomo, A. J. Lees, S. E. Daniel, C. Colosimo, G. Wenning, and N. Quinn. Accuracy of clinical diagnosis

of progressive supranuclear palsy. *Movement disorders*, 19(2): 181–189, 2004.

E. Oz and H. Kaya. Support vector machines for quality control of DNA sequencing. *Journal of Inequalities and Applications*, 2013(1):85, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

S. Peng, Y. Ma, P. G. Spetsieris, P. Mattis, A. Feigin, V. Dhawan, and D. Eidelberg. Characterization of disease-related covariance topographies with SSMPCA toolbox: Effects of spatial normalization and PET scanners. *Human brain mapping*, 35 (5):1801–1814, 2014.

P. Perner. Improving the accuracy of decision tree induction by feature pre-selection. *Applied Artificial Intelligence*, 15(8): 747–760, 2001.

K. Polat and S. Güneş. A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36(2, Part 1):1587–1592, 2009. doi> 10.1016/j.eswa.2007.11.051

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA, 1993.

J. R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996a.

J. R. Quinlan. Learning decision tree classifiers. *ACM Computing Surveys*, 28(1):71–72, 1996b.

P. E. Rauber, R. R. O. d. Silva, S. Feringa, M. E. Celebi, A. X. Falcão, and A. C. Telea. Interactive Image Feature Selection Aided by Dimensionality Reduction. In *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2015.

L. Rokach and O. Maimon. Classification trees. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery*

Handbook, pages 149–174. Springer US, 2010. ISBN 978-0-387-09822-7.

R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, and E. Bullmore. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral Cortex*, 15(9):1332–1342, 2005.

A. Sato and K. Yamada. Generalized learning vector quantization. *Advances in neural information processing systems*, pages 423–429, 1996.

E. F. Schisterman, N. J. Perkins, A. Liu, and H. Bondell. Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, 16(1): 73–81, 2005.

P. Schneider, M. Biehl, and B. Hammer. Relevance matrices in LVQ. In *Proc. Of European Symposium on Artificial Neural Networks (ESANN 2007)*, pages 37–42. d-side publishing, 2007.

P. Schneider, M. Biehl, and B. Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009.

J. M. Shulman and P. L. De Jager. Evidence for a common pathway linking neurodegenerative diseases. *Nature genetics*, 41(12):1261–1262, 2009.

D. H. Silverman. Brain 18F-FDG PET in the diagnosis of neurodegenerative dementias: comparison with perfusion SPECT and with clinical evaluations lacking nuclear imaging. *Journal of Nuclear Medicine*, 45(4):594–607, 2004.

P. G. Spetsieris, V. Dhawan, and D. Eidelberg. Three-fold cross-validation of parkinsonian brain patterns. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 2906–2909, 2010. doi> 10.1109/IEMBS.2010.5626327

P. G. Spetsieris and D. Eidelberg. Scaled subprofile modeling of resting state imaging data in Parkinson's disease: Methodological issues. *NeuroImage*, 54(4):2899–2914, 2011. doi> 10.1016/J.NEUROIMAGE.2010.10.025

P. G. Spetsieris, Y. Ma, V. Dhawan, and D. Eidelberg. Differential diagnosis of parkinsonian syndromes using PCA-based

functional imaging features. *NeuroImage*, 45(4):1241–1252, 2009. doi> 10.1016/j.neuroimage.2008.12.063

G. Stiglic, S. Kocbek, I. Pernek, and P. Kokol. Comprehensive decision tree models in bioinformatics. *PloS one*, 7(3):e33812, 2012.

C. C. Tang, K. L. Poston, T. Eckert, A. Feigin, S. Frucht, M. Gudesblatt, V. Dhawan, M. Lesser, J.-P. Vonsattel, S. Fahn, and D. Eidelberg. Differential diagnosis of parkinsonism: a metabolic imaging study using pattern analysis. *The Lancet Neurology*, 9(2):149–158, 2010. doi> 10.1016/s1474-4422(10)70002-8

L. K. Teune, A. L. Bartels, B. M. de Jong, A. T. Willemsen, S. A. Eshuis, J. J. de Vries, J. C. van Oostrom, and K. L. Leenders. Typical cerebral metabolic patterns in neurodegenerative brain diseases. *Movement Disorders*, 25(14):2395–2404, 2010.

L. K. Teune, D. Mudali, R. J. Renken, B. M. D. Jong, M. Segbers, J. B. T. M. Roerdink, R. A. Dierckx, and K. L. Leenders. Glucose IMaging in ParkinsonismS. In *16th International Congress of Parkinson's Disease and Movement Disorders, Dublin, Ireland June 17-21*, 2012. Abstract # 783.

L. K. Teune, R. J. Renken, D. Mudali, B. M. D. Jong, R. A. Dierckx, J. B. T. M. Roerdink, and K. L. Leenders. Validation of parkinsonian disease-related metabolic brain patterns. *Movement Disorders*, 28(4):547–551, 2013. doi> 10.1002/mds.25361

B. Thompson. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55:525–534, 1995.

D. Townsend. Physical principles and technology of clinical PET imaging. *Annals-Academy of Medicine Singapore*, 33(2):133–145, 2004.

M. Ture, F. Tokatli, and I. Kurt. Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications*, 36(2, Part 1):2017–2026, 2009. doi> 10.1016/j.eswa.2007.12.002

M. P. Van Den Heuvel and H. E. H. Pol. Exploring the brain network: a review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, 2010.

K. Van Laere, C. Casteels, L. De Ceuninck, B. Vanbilloen, A. Maes, L. Mortelmans, W. Vandenberghe, A. Verbruggen, and R. Dom. Dual-tracer dopamine transporter and perfusion SPECT in differential diagnosis of parkinsonism using template-based discriminant analysis. *Journal of Nuclear Medicine*, 47(3):384–392, 2006.

M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 110–119. ACM, 2000.

M. A. Westenberg and J. B. T. M. Roerdink. Mixed-method identifications. In J. M. H. Du Buf and M. M. Bayer, editors, *Automatic Diatom Identification*, volume 51 of *Series in Machine Perception and Artificial Intelligence*, chapter 12, pages 245–257. World Scientific Publishing Co., Singapore, 2002.

P. Wu, J. Wang, S. Peng, Y. Ma, H. Zhang, Y. Guan, and C. Zuo. Metabolic brain network in the Chinese patients with Parkinson's disease based on 18 F-FDG PET imaging. *Parkinsonism & Related Disorders*, 19(6):622–627, 2013.

L. Yeo, N. Adlard, M. Biehl, M. Juarez, T. Smallie, M. Snow, C. Buckley, K. Raza, A. Filer, and D. Scheel-Toellner. Expression of chemokines CXCL4 and CXCL7 by synovial macrophages defines an early stage of rheumatoid arthritis. *Annals of the rheumatic diseases*, pages annrheumdis–2014, 2015.

W. Youden. An index for rating diagnostic tests. *Cancer*, 3: 32–35, 1950.

M. Yun, W. Kim, N. Alnafisi, L. Lacorte, S. Jang, and A. Alavi. 18F-FDG PET in characterizing adrenal lesions detected on CT or MRI. *Journal of Nuclear Medicine*, 42(12):1795–1799, 2001.

X. Zhang. Using class-center vectors to build support vector machines. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop.*, pages 3–11, Aug 1999. doi> 10.1109/NNSP.1999.788117

Y. Zheng, H. Suematsu, T. Itoh, R. Fujimaki, S. Morinaga, and Y. Kawahara. Scatterplot layout for high-dimensional data visualization. *Journal of Visualization*, 18(1):111–119, 2014.

PUBLICATIONS

JOURNAL PAPERS

D. Mudali, L. K. Teune, R. J. Renken, K. L. Leenders, and J. B. T. M. Roerdink. "Classification of Parkinsonian Syndromes from FDG-PET Brain Data Using Decision Trees with SSM/PCA Features", Computational and Mathematical Methods in Medicine, Article ID 136921:1–10, 2015. DOI: http://dx.doi.org/10.1155/2015/136921.

L. K. Teune, R. J. Renken, D. Mudali, B. M. De Jong, R. A. Dierckx, J. B. T. M. Roerdink, and K. L. Leenders. "Validation of parkinsonian disease-related metabolic brain patterns. Movement Disorders", 28(4):547–551, 2013. DOI: http://dx.doi.org/10.1002/mds.25361.

PEER-REVIEWED CONFERENCE PAPERS

D. Mudali, M. Biehl, K. L. Leenders, and J. B. T. M. Roerdink. LVQ and SVM Classification of FDG-PET Brain Data. Advances in Self-Organizing Maps and Learning Vector Quantization. Proc. WSOM 2016, 11th Workshop on Self-Organizing Maps. E. Merényi et al. (eds.), Advances in Intelligent Systems and Computing 428, Springer International Publishing Switzerland, 2016, DOI = http://dx.doi.org/10.1007/978-3-319-28518-4_18.

D. Mudali, L. K. Teune, R. J. Renken, K. L. Leenders and J. B. T. M. Roerdink. Comparison of Decision Tree and Stepwise Regression Methods in Classification of FDG-PET Brain Data using SSM/PCA Features. 8th International Conference on Advanced Computational Intelligence, ICACI, Thailand, February 14-16, 2016.

POSTERS AND ABSTRACTS

D. Mudali, L. K. Teune, R. J. Renken, K. L. Leenders, and J. B. T. M. Roerdink. "Comparison of decision tree and stepwise regression methods in classification of FDG-PET data". In Third European Conference on Clinical Neuroimaging, March 31-April 1, Lille, France. Page 16, 2014. Abstract.

D. Mudali, L. K. Teune, R. J. Renken, K. L. Leenders, and J. B. T. M. Roerdink. "Decision Tree Classification of FDG-PET Data to Predict Neurodegenerative Diseases". ICT-OPEN 2012 ASCI, October 22, 2012, Rotterdam (poster).

L. K. Teune, D. Mudali, R. J. Renken, B. M. De Jong, M. Segbers, J. B. T. M. Roerdink , R. A. Dierckx, and K. L. Leenders. Glucose IMaging in ParkinsonismS. In 16th International Congress of Parkinson's Disease and Movement Disorders, Dublin, Ireland June 17-21. 2012. Abstract

SAMENVATTING

VOORSPELLING VAN NEURODEGENERATIEVE AANDOENINGEN OP BASIS VAN FUNCTIONELE HERSENAFBEELDING

Neurodegeneratieve aandoeningen leiden tot uitdagende problemen voor maatschappijen met een hoge levensverwachting. Als er niets aan gedaan wordt zullen deze aandoeningen effect blijven hebben op ouderen en het sterftecijfer blijven verhogen. Aangezien ze een traag verloop hebben zijn ze lastig in een vroeg stadium te herkennen. Bovendien vertonen patiënten met deze aandoeningen allemaal vergelijkbare ziekteverschijnselen, wat ze lastig te onderscheiden maakt.

In dit proefschrift was het doel technieken te onderzoeken om biomarkers te verkrijgen uit hersenbeelden die gemaakt zijn door middel van Positron Emissie Tomografie (FDG-PET), om langs die weg neurodegeneratieve aandoeningen te kunnen voorspellen en classificeren, in het bijzonder Parkinson-achtige aandoeningen. We hebben principale component-analyse (PCA) gecombineerd met het geschaalde subprofielmodel (SSM) om kenmerken uit de hersenbeelden te halen die geschikt zijn voor classificatie. De resulterende classificatiemethoden zijn gevalideerd.

Hoofdstuk 1 geeft achtergrondinformatie over neurodegeneratieve aandoeningen en technieken voor het in beeld brengen van het brein. In hoofdstuk 2 beginnen we onze studie naar de classificatie van Parkinson-achtige aandoeningen met "decision trees" (besluitbomen), aangezien deze laatste vrij intuïtief zijn. Kenmerken in de vorm van subject-scores zijn uit de hersenbeelden gehaald met de SSM/PCA methode. Deze kenmerken zijn vervolgens doorgegeven aan het C4.5 algoritme om een besluitboom te trainen. Deze is vervolgens gevalideerd met de "leave-one-out" methode. De resultaten laten zien dat een besluitboom onderscheid kan maken tussen gezonde proefpersonen en MSA (multiple system atrophy) of PSP (progressive supranuclear palsy) patiënten, maar meer moeite heeft met PD (Parkinson's disease) gevallen, hoewel de resultaten van de besluitboom verbeteren door alleen de meest relevante ken-

merken te gebruiken. We concluderen dat PD patiënten lastig te onderscheiden zijn van de gezonde controlegroep (zeker als de aandoening nog in een vroeg stadium is), en dat dit komt omdat de FDG-PET activatiepatronen voor deze twee groepen vergelijkbaar zijn. Bovendien was de hoeveelheid data die tot onze beschikking stond in dit hoofdstuk te klein om betere resultaten te bereiken. Paarsgewijze vergelijkingen van de ziektegroepen (zonder de controlegroep) leverde betere classificatieresultaten op. Andere classificatie-algoritmen, zoals lineaire discriminant-analyse (LDA), "nearest neighbors", classificatie-en-regressiebomen (CART), "random forests", enz. zijn ook uitgeprobeerd, maar lieten geen uitzonderlijk goede resultaten zien. De besluitbomen zijn bovendien van belang voor het visualiseren van de classificatieresultaten, en geven hiermee inzicht in de distributie van de kenmerken.

In hoofdstuk 3 vergelijken we besluitbomen met stapsgewijze regressie (SR), welke probeert enkele "goede" PCA componenten lineair te combineren. De SR methode was effectiever dan de besluitbomen voor het classificeren van Parkinson-achtige aandoeningen. Dit komt omdat de SR-methode de beste kenmerken combineert tot één robuust kenmerk voor het classificeren van de Parkinson-achtige aandoeningen. De besluitbomen daarentegen gebruiken de individuele kenmerken afzonderlijk. Interessant genoeg zien we dat de accuratesse even goed is als hetzelfde robuuste kenmerk werd gebruikt voor het maken van een besluitboom. Het is dus zeker mogelijk de twee methoden te combineren waarbij de kenmerken door de SR-methode geleverd worden en de classificatiemethode een besluitboom gebruikt. Het voordeel van het gebruik van een besluitboom is dat die, in tegenstelling tot de stapsgewijze regressiemethode, in staat is meer dan twee groepen te onderscheiden.

De besluitboom-methode was onze eerste keuze voor het classificeren van neurodegeneratieve aandoeningen vanwege het feit dat besluitbomen makkelijk te begrijpen zijn. Echter, vanwege de niet erg bevredigende resultaten in de eerdere hoofdstukken, onderzoeken we in hoofdstuk 4 enkele andere classificatie-algoritmen. In dit hoofdstuk passen we "Generalized Matrix Learning Vector Quantization" (GMLVQ) en "Support Vector Machine" (SVM) classificatie-algoritmen toe op data van Parkinson-achtige aandoeningen in de hoop hiermee betere classificatieresultaten te behalen. Net als eerder gebruiken we de SSM/PCA-methode om geschikte kenmerken voor de classificatie te verkrijgen en geven we deze door aan de GMLVQ- en

SVM-methoden. Uit de resultaten blijkt dat zowel GMLVQ en SVM beter zijn dan besluitbomen in het classificeren van in een vroeg stadium optredende Parkinson-achtige aandoeningen. SVM is praktisch even goed als GMLVQ in binaire classificatie, maar de "Oppervlakte Onder de Curve"-maten laten zien dat GMLVQ beter is. Als het gaat om het onderscheiden van meer (dan twee) groepen is GMLVQ beter dan SVM. In tegenstelling tot SVM zijn de GMLVQ resultaten makkelijk te interpreteren in termen van diagonale en neven-diagonale matrices.

We passen de besluitboom-, QMLVQ- en SVM-methoden ook toe op een later verkregen dataset met een groter aantal scans van patiënten in een verder gevorderd stadium van PD. In hoofdstuk 5 vergelijken we de mate waarin de verschillende methoden de verder gevorderde PD-gevallen van de gezonde controlegroep kunnen onderscheiden. De "leave-one-out cross validation" resultaten voor de besluitbomen zijn hier veel beter dan de resultaten in hoofdstuk 2. De GMLVQ- en SVM-methoden zijn veel beter in staat dan de besluitbomen om de vroege PD-gevallen te onderscheiden van de gezonde controlegroep. Aan de andere kant zijn GMLVQ en besluitbomen intuïtiever dan SVM. Ze kunnen beide erg grote datasets aan. We komen tot de conclusie dat trainen (en testen) met een grotere dataset waar ook verder gevorderde PD-gevallen in voorkomen veel betere resultaten geeft dan trainen met een kleinere dataset waar alleen vroege PD-gevallen in zitten. Grotere datasets leiden dus tot betere classificatie van neurodegeneratieve aandoeningen. Alle classificatiemethoden in dit proefschrift werken goed met data van patiënten in een verder gevorderd stadium van de aandoening. We komen zo tot de eindconclusie dat GMLVQ en besluitbomen interessant zijn voor verder onderzoek naar de classificatie en voorspelling van neurodegeneratieve aandoeningen.

ACKNOWLEDGEMENTS

The completion of this thesis has been possible with the contribution and support of many people. First and foremost I would like to give glory to God almighty who held me all through the season.

My supervisor and promotor Prof. dr. Jos B.T.M. Roerdink, you are exceptional, I am so grateful to you. Thank you for this marvelous opportunity to do my PhD studies under your guidance, mentorship and support. I have learned and grown intellectually through your productive and innovative ideas. I will forever be indebted to you for accepting me from the first e-mail I wrote to you expressing my interest to work with you. Thank you for your positivity and belief in me, more so the encouragement. Dank je.

To my second supervisor Prof. Michael Biehl, thank you for the valuable support especially towards the end of the journey. You always provided detailed feedback and critical comments. Thank you for introducing me to LVQ.

Special thanks go to the members of the reading committee i.e., Prof. B. M. ter Haar Romeny, Prof. N. M. Maurits and Prof. A. C. Telea for their constructive and informative comments.

I am grateful to Dr. John A. Quinn for providing a research space and enabling me to present my work in his research group. Thank you for the contribution to my success.

To the people at the NIC and UMCG who made this work possible through provision of brain image data and productive discussions I am grateful. Laura Teune, thanks for being available especially when I needed to consult you. I would also like to thank Remco Renken and Klaus Leenders for the constructive and critical comments that caused me to think a lot to come up with solutions. Thanks to Sanne too.

To the NUFFIC (NFP) program, my sponsor, I am so grateful for the generous provision that made this work possible. Again thank you for the remarkable opportunity. Particularly, Erik Haarbrink thank you for making my stay in Groningen worthwhile. My special appreciation goes out to Gonny Lakerveld for the special support she provided to me when I lost my mother.

Thank you very much for your time and care which uplifted me to continue pursuing my studies. Also, you always made my stay in Groningen comfortable. I am generally in appreciation of the NUFFIC program for providing such exceptional opportunities to study in the Netherlands.

To my colleagues in the SVCG group, thank you for making the learning environment conducive with the discussions. Jasper thank you so much, for you were always willing to help with all research related issues. Alessandro my first officemate thanks for ushering me in the research environment and for your initial guidance and help during the beginning of my PhD journey. I am also grateful to have met Bilkis, thank you for the guidance and help. In the same manner I thank the rest of the group members; Yun, Moritz, Ozan, Maarten, Andre, Matthew, David, Cong and Chengtao. I would also like to thank Dr. Henk Bekker for the discussions at academic conferences.

The PhD students from the intelligent systems group who also made my journey smooth: Ugo M thanks for the technical help always and Ernest M thanks for your help with LVQ.

To the administrative staff at the Johann Bernoulli Institute for Mathematics and Computer Science, particularly Desiree Hansen, thanks for always making me a happy researcher, keep up with the spirit of being jolly. Also my sincere gratitude goes to Ineke Schelhaas and Esmee Elshof. Further more, to the PhD coordinator Janieta de Jong-Schlukebir, thank you for the uplifting words. Generally, I appreciate every member of the Johann Bernoulli Institute for Mathematics and Computer Science.

I am grateful to all my friends who made life worth living in Groningen, the Netherlands, i.e., Tendero, Sampson, Susan, Fred Noah, Prossy, Yany, Victor N, Dajo, Antonia, Louisa, Uchenna, Peace, Rosecarmen, Johnson, Annelies, Anne K, Shereen and all others who have been there for me one way or another. Dichic thank you for caring.

Finally, to my siblings and father, a special warm appreciation for your unconditional support and love. To my mother in a nutshell, 'you are my guardian ANGEL'.

CURRICULUM VITAE

The author studied Computer Science at Mbarara University of Science and Technology, Uganda, from 2001-2004, where she obtained her BSc degree. Her final project was entitled "Pharmacy monitoring system". From 2004-2006 she was enrolled in the master program Computer Science at Makerere University, Kampala, Uganda. Her MSc thesis was entitled "A web based medical data and image repository". From 2007-2009 she was a lecturer at the Institute of Computer Science, Mbarara University of Science and Technology.

She was a PhD student at the Johann Bernoulli Institute for Mathematics and Computer Science of the University of Groningen from 2010-2014. Since 2014 until now she has been a research fellow at the Artificial Intelligence Group of Makerere University. Since 2015 she has returned as lecturer to Mbarara University of Science and Technology.

Her research interests include machine learning, medical image analysis, pattern recognition, scientific visualization, and bioinformatics.

COLOPHON

This thesis was typeset with LaTeX 2ε using Robert Slimbach's *Minion Pro* type face. The style of this thesis is based on André Miede's excellent Classic Thesis Style.