Heinrich-Heine Universität Düsseldorf

Data Representation and Classification of Alzheimer's Disease

Master Thesis

Institute of Neuroscience and Medicine (INM-7)

Research Centre Jülich, Jülich, Germany

Submitted by

Jean-Philippe Kröll, B. Sc.

Martriculation number : 2094289

On the 21st October 2019

In the Master Course

Translational Neuroscience

Faculty of Medicine

First advisor: Univ.-Prof. Dr. Simon B. Eickhoff

Second advisor: Dr. Kaustubh Patil

EIDESSTATTLICHE VERSICHERUNG

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Düsseldorf, 21.10.2019

Ort, Datum

Jean-Philippe Kröll,

Name (Unterschrift)

Table of Content

List of Abbreviations

AD = Alzheimer’s Disease

AMAP = Adaptive Maximum A Posterior

C = Capacity

CA = Cornus Ammonis

CDR = Clinical Dementia Rating

CV = Cross Validation

FWHM = Full Width at Half Maximum

GS = Geodisic Shooting

HC = Healthy Controls

LAS = Local Adaptive Segmentation

MCI = Mild Cognitive Impairment

MFS = Magnetic Field Strength

MMSE = Mini Mental State Examination

MRI = Magnetic Resonance Imaging

NFM= Non-selected Feature Model

OM = Original Model

PCA = Posterior Cortical Atrophy

PCC = Pearson’s Correlation Coefficient

PCC = Posterior Cingulate Cortex

PVE = Partial Volume Estimation

SANLM = spatial-adaptive non-local means

SFM = Selected Feature Model

SNR = Signal-to-Noise Ratio

SVM = Support Vector Machine

TPM = Tissue Probability Map

Abstract

Application of machine learning algorithms to information of magnetic resonance imaging (MRI) is a widespread approach to differentiate Alzheimer's disease (AD) patients and healthy controls (HC). Since a variety of brain representations are used by different studies, it is necessary that the influence of the chosen brain atlas on the model performance is investigated. Therefore, the goal is to analyse the effect which is caused by varying granularity of an atlas. In addition, to find acceptance in the medical community, the model must be able to identify biologically relevant regions. Thereby, it can be ensured that the model will reliably identify patients in future applications and is not based on sample-specific characteristics. For this reason, the regions selected for classification by support vector machine (SVM) to differentiate AD vs HC are analysed. Lastly, features that are not selected by a given model are generally disregarded. Since those features could potentially still contain relevant information, they are examined in this study. Different granularities of the Schaefer atlas, with parcellations ranging from 173 to 1273 parcels, were used to extract features from structural images of AD patients and healthy controls. Subsequently, SVM classifiers were trained on the features derived from the different parcellations and their influence was evaluated based on the performance of the resulting model. Biological relevance of the selected features was verified by confirming their role in AD with current literature. Non-selected features were singled out and used to train a non-selected feature model (NFM). Relevance of the non-selected features was evaluated based on performance of the NFM. Evaluation of the obtained accuracies showed that the granularity of the atlas affects the model performance on 1.5 Tesla images of AD patients and HC. Accuracies ranged from 87% for the 173 parcel parcellation, to 83% for the 1273 parcel parcellation. Classification of 3 Tesla images was not significantly affected, with all models achieving accuracies around 91%. Biological relevance of the selected features could be confirmed by literature, although it was evident that not all relevant regions were included in the model. Examination of the NFM revealed that a model based on non-selected features could still classify AD vs HC with an accuracy of 76%. The findings suggest that future atlas-based approaches should pay more attention to the effect of the selected atlas. In addition, the ability of SVM to select biologically relevant regions supports its implementation for diagnosis of AD in the clinic. Lastly, the results indicate that investigation of non-selected features could provide additional insight into the relevance of certain regions for the studied disease.

Introduction

Alzheimer's disease

Alzheimer´s disease is the most common form of dementia and affects around 30 million people worldwide [1]. The World Health Organization lists Alzheimer and other dementias as the third leading cause of death in high income countries [2]. Symptoms of the disease vary among patients, but the initial symptom is usually impaired memory[3]. As the disease progresses, individuals often show difficulties with speech, attention and judgement[3]. The pace at which cognitive abilities decline differs, but the average life expectancy from the time point of diagnosis is 3 to 10 years, depending on the patient's age at that time[4]. Currently, no treatment or prevention is known. The underlying mechanisms are poorly understood, but with progression of the disease, brain volume is lost continuously. Therefore researchers agree, that intervention at the earliest possible stage is crucial to successfully help patients. Since it is known, that changes in the brain occur years before first clinical symptoms manifest, one of the main objectives of Alzheimer's research is the early detection of the disease. Therefore, a lot of research has focused on developing biomarkers which can categorize mild cognitive impairment (MCI) patients into converters and non-converters. MCI is defined as a decline in cognitive abilities, which exceeds the expected decline for an individual's age and education, yet doesn't obstruct the individuals daily life[5]. MCI is widely accepted as an early stage of dementia, especially of AD. Nevertheless, not all MCI patients progress to dementia. The annual conversion rate is estimated to be around 7%, emphasizing the need for biomarkers which differentiate converters from non-converters[6].

Structural MRI as biomarker for Alzheimer's disease

Biomarkers are measurable indicators of a biological state. In Alzheimer's research, it is a constant endeavour to find a biomarker which can reliably track, or better, predict the onset of AD. Currently, structural images are investigated for their usefulness as a biomarker for AD. The structure of the brain is subject to alterations throughout the disease[7,8]. Structural images represent a non-invasive opportunity to track those alterations. The degree of brain atrophy shows a strong correlation with the severity of symptoms and gives insights into the progression from normal cognition to dementia[9]. More so, serial imaging studies have shown, that increased atrophy rates in cognitively normal subjects are a strong indicator for the development of AD[10]. This makes structural images a suitable candidate for early prediction of AD onset. Although there exist other biomarkers, like CSF-tau, which can detect the disease even sooner, the use of structural MRIs has one major advantage: Atrophy measured by structural MRI remains highly correlated with cognitive decline, even in later stages of disease progression[11]. In contrast, CSF-tau levels stay relatively stable during the later stages of the disease and are therefore not suited for tracking AD in more progressed states[12]. Atrophy rates are also useful to investigate the transition from the prodromal stage of AD, mild cognitive impairment, to dementia. Several studies showed that structural MRI, especially when combined with other markers, is well suited to predict conversion from MCI to AD[13,14].

Support Vector Machine

To analyse the enormous amount of data provided by structural images, learning algorithms are used. In this study, Support Vector Machine (SVM) was implemented to classify AD vs healthy controls. SVM was chosen because it has been shown to perform well for the classification of AD[15–17]. Classifiers learn to differentiate two classes by training on labelled data. In this case, features extracted from the brain images of AD and HC subjects were used to train the model.

Data representation

Working with brain data, one of the first obstacles is to decide which brain atlas to use. Since there exist no guidelines, it has become increasingly difficult to choose among the ever-growing number of available atlases. This has led to the fact, that similar problems have been investigated with a variety of anatomical representations. For example in the case of classification of Alzheimer's disease, researchers have used the Automatic Anatomic Labeling atlas, the LONI Probabilistic Brain Atlas, or multi-atlas approaches [18–20]. Disadvantageously, there is no agreement on the impact which the selected brain atlas has on the performance of the classifier. Machine learning approaches are highly influenced by the choice of the atlas since the atlas-based parcels constitute the features used by the algorithm. In fact, each atlas divides the brain into different numbers of regions, with varying borders and sizes. When comparing classification performance of different models, much emphasis is usually put on the implemented algorithm. Rarely, the effect of the chosen brain atlas is acknowledged. To enlighten this effect, we investigated the influence of the chosen data representation on the classification of Alzheimer's disease. Although this has been studied before (Long et al., 2018 and Ota et al., 2014), this study is, to our knowledge, the first to use such a large cohort and investigate the influence on different magnetic field strengths separately[15,21].

Predictive regions

Furthermore, we wanted to investigate if the predictors of an SVM based classifier would also be biologically relevant for AD. Identification of disease relevant regions is crucial for implementation for diagnosis in clinical routine. Since regions affected in AD are well known, the selected features of the classifier in this study can easily be verified. We therefore hoped to confirm the suitability of implementing SVM classifiers for disease pattern identification in clinical approaches.

Feature selection

In neuroimaging studies, one is usually confronted with an enormous amount of features. Especially in voxel-based approaches, the number of features (corresponding to the voxels of each image) vastly exceeds the number of samples (image of each subject). Although an atlas-based approach was used in this study, which reduces the amount of features to a maximum of 1273 regions, the dimensionality of the features is still tremendous. Since an excess of features will increase computational cost, and even worse, can impair the classifiers performance, most studies perform feature selection. Feature selection describes the process of selecting a subset of features, which are most relevant to build the given model[22]. The assumption behind this approach is that the data contains features which are irrelevant for the given task or redundant in relation to other (relevant) correlated features. In our case, each feature contains information about the brain, but not every region is relevant for Alzheimer's disease. A further advantage of feature selection is, that it improves interpretability of the model. By investigating the selected features of a model, it is possible to get a comprehensive insight into the basis on which the classification is done. This is especially important if the goal is to apply the model to clinical settings. To ensure that the classifier will accurately differentiate patients from healthy subjects in the future, it has to be ensured, that the decision of the algorithm is based on universally applicable features. In this case, we anticipated to be able to trace back the selected regions to brain areas known to be affected in AD.

Value of non-selected features

To achieve the best performance, it is often not useful to select the most relevant features, as they may contain redundant information[23]. Rather, a subset of features with complementary informative value will result in the highest accuracy. By applying feature selection, non-relevant and redundant features are excluded to improve performance. Only a subset of all available features are selected for classification. While many studies have investigated the features selected by such an approach, few have considered the value of features which are not selected. Especially in a medical context, those features may still provide insight into the mechanisms of the disease. Therefore, during this study, we wanted to investigate if non-selected features still contained relevant information. For this purpose, a model was built which contained only features that were removed by feature selection in the original model. Since feature selection can also remove correlated features that are redundant, also features which were correlated to the selected features were excluded. Thereby it was ensured that the performance of the non-selected feature model would be solely based on features unutilized by the original model. Subsequently, the model was evaluated on its ability to classify AD patients vs HC. Additionally, we wanted to see whether the non-selected feature model would outperform the original features in classification of MCI patients. We hypothesized, that specific regions could play an important role in earlier stages of the disease, but be less important in comparison to highly atrophied regions in later stages.

Methods

Subjects

For the analysis, structural T1-weighted images of 449 subjects (149 Alzheimer Patients and 300 healthy controls) where taken from the Alzheimer's disease Neuroimaging Initiative (Table 1). Exclusion criteria were neurological diseases other than AD, history of head trauma with following neurological impairment or abnormal brain structure. Diagnosis of AD was based on Mini-Mental State Examination (score between 20-26), Clinical Dementia Rating (0.5 or 1) and NINCDS/ARDRA criteria[24] . Since the main interest of this study was the early diagnosis of Alzheimer's disease, only the scans from the first visit of each participant were used. To ensure high signal to noise ratio, only 3 Tesla images were included.

Since not all hospitals work with MRI scanners with identical magnetic field strength, we also wanted to investigate if images with different magnetic field strengths would be influenced differently by the atlases. Therefore, a second group was formed, which contained only 1.5 Tesla images. Images of 372 participants (166 Alzheimer patients and 206 controls) were taken from ADNI and used for comparison (Table 1).

Additionally, structural T1-weighted images of 413 MCI subjects (138 early MCI patients, 137 late MCI patients and 138 MCI-Converters) where taken from ADNI (Table 2). Only 3T images were included. Diagnosis of mild cognitive impairment was based on subjective memory complaints, Mini-Mental State Examination (score between 24-30), Clinical Dementia Rating (at least 0.5) and sufficiently preserved cognition, such that a diagnosis of AD cannot be made at the day of screening. Early MCI subjects and late MCI subjects were subdivided based on their score on the Logical Memory II subscale of the Wechsler Memory Scale. Ranges were adjusted according to years of education and early MCI patients had to score higher than late MCI patients. Only subjects who transitioned from MCI to AD were labelled as MCI-Converters. This included both, subjects originally assigned to early MCI and subjects originally assigned to late MCI. Out of the Non-Converter groups in this study, neither early MCI, nor late MCI subjects included patients who converted to AD up to this day.

Table 1
Subject group characteristics of AD patients and healthy controls

	3T images		1.5T images	
	Controls	AD	Controls	AD
Number	306	153	206	166
Male/Female	154/229	88/65	107/99	88/78
Age (years)	75.2 ± 5	73.6 ± 4.7	75.3 ± 4.6	77.3 ± 6.3
CDR (Score)	0	1	0	1
MMSE (Score)	28.8 ± 2.6	22.3 ± 6.5	29.1 ± 1	22.8 ± 5.3
Education (years)	15.7 ± 2.7	16.4 ± 2.6	16.2 ± 2.8	13.6 ± 3.1

Mean ± Standard Deviation. AD – Alzheimer's Disease, CDR – Clinical Dementia Rating, MMSE – Mini Mental State Examination

Table 2
Subject group characteristics of MCI patients

	eMCI	lMCI	Converter
Number	138	137	138
Male/Female	75/63	83/54	78/60
Age (years)	72.7 ± 9.2	77.8 ± 4.6	74.8 ± 6.8
CDR (Score)	0.5	0.5	0.5
MMSE (Score)	28.1 ± 2.8	26.5 ± 4.9	27.4 ± 1.8
Education (years)	15.8 ± 2.7	15.9 ± 3.0	15.9 ± 2.6

Mean ± Standard Deviation. eMCI – early Mild Cognitive Impairment, lMCI – late Mild Cognitive Impairment, CDR – Clinical Dementia Rating, MMSE – Mini Mental State Examination

Schaefer atlas

The brain representation used in this study is the Schaefer parcellation of the human cerebral cortex[25]. Cortical parcels were estimated based on resting state fMRI from 1489 subjects. While previous publications relied on either global similarity or local gradient methods, Schaefer et al. (2018) combined both approaches[25]. A gradient-weighted Markov Random Field was applied, with three competing terms. The first term, representing the global similarity approach, assigns regions with similar fMRI time courses to the same label. The second term, which corresponds to the local gradient approach, encourages adjacent parcels with high gradients between them to have different labels. The third term constrains brain areas constituting a parcel to be close to the centre of that parcel, to account for long range resting-state functional connectivity. The Schaefer parcellation was chosen, because it is available in different resolutions, ranging from 100 to 1200 parcels (for example see figure 1). This allowed us to investigate the influence of varying regions, with different sizes and borders. In previous studies, multiple atlases were used to study this effect[15,21]. This comes with the drawback that one has to assume that the results are not further influenced by the different methods used to build the atlases. Using only the Schafer atlas allowed us to eliminate this bias and directly compare the effect of granularity. Finally, the atlas is based on a large cohort and was shown to be homogeneous across different acquisition protocols.

In this study, the Schaefer atlas was complemented by subcortical structures from the Brainnetome atlas[26] and the cerebellum from the publication of Buckner et al. (2011)[27].

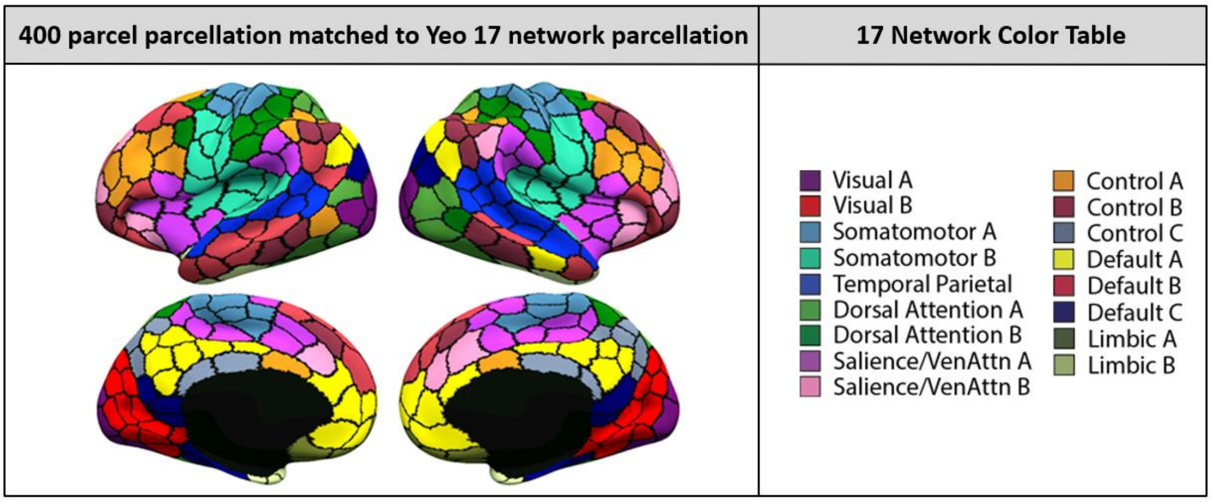

Figure 1. Visualization of the 400 parcel parcellation in fslr32k space, colored to match Yeo 17 network parcellation[28]. (Retrieved November 10th, 2019, from https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal)

Linear support vector classification

For this study, Support Vector Machine (SVM) was chosen to classify Alzheimer's patients vs healthy controls. SVM describes a class of linear and non-linear supervised learning models, which can be used for classification and regression. In this study, SVM with a linear kernel was used. The ultimate goal of SVM is the separation of different classes by a hyperplane, which results in the least misclassifications. To achieve this, SVM finds the separating hyperplane with maximal distance to the closest data points (Figure 2). The distance to the closest data points is also called the margin of the classifier. Maximizing the margin ensures that the decision boundary is as far away as possible from the most uncertain data points of each class. Thereby the classifier will perform well on new data, even when confronted with outliers of each class. The separator is constructed using only on a few data points, also known as the support vectors. SVM assigns new data points to a class, based on which side of the separator they fall. Since the support vectors determine the position of the separator, SVM is only dependent on a subset of all data points [29]. SVM is therefore computational efficient and generally results in lower prediction time.

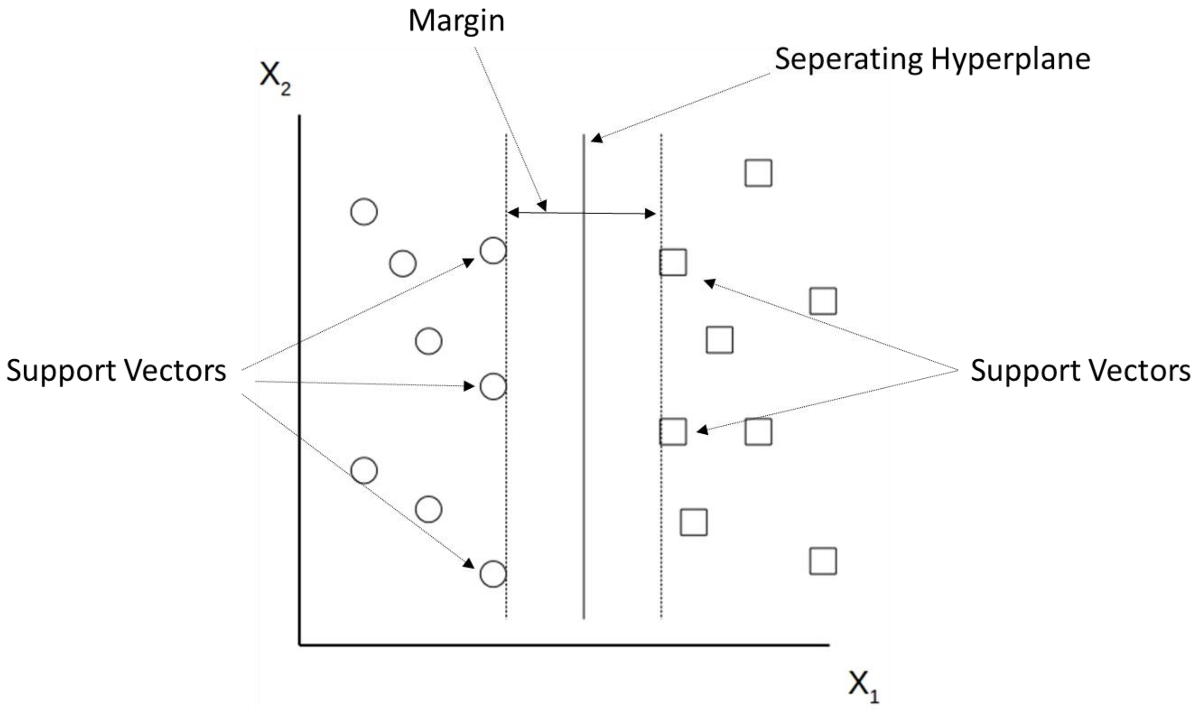

Figure 2. Depiction of a SVM. Different symbols (circles and squares) indicate the two classes which are to be separated. For simplicity, only two dimensions are shown, in which case the classes can be separated by a line. Note that only the most outer points of each class (the support vectors) determine the position of the separator.

Generalization of a model

A major goal of any machine learning approach is to create a model which performs well on unseen data. This is referred to as the generalization ability, which itself is highly influenced by the complexity of the model. Here, the model complexity refers especially to the amount of features included in a given model. By integrating more features, the model becomes more complex. Choosing the optimal model complexity is subject to the Bias-variance trade-off[30]. Bias describes the difference between the predicted value and the actual truth. A model with high bias fails to capture the underlying pattern of the dataset because it overgeneralizes. In other words, the model is too simple and underfits the data. This results in high prediction error for both, training and test sets. On the other hand, variance describes the model's sensitivity to noise of the dataset. While any algorithm, given sufficient training, is capable of capturing every detail of the trainings set, this does not translate to good performance on new data. Quite the opposite – by capturing every detail, the model will capture noise inherent to the trainings set and therefore be unable to generalize when confronted with unseen data points. In other words, the model is too complex and overfits the data. High variance can result in perfect accuracy for the training set, but very poor performance on the test data. It becomes evident, that the trade-off between bias and variance is simultaneously a trade-off between over- and under-fitting. Since one cannot be reduced without increasing the other, optimizing a model includes finding the level of model complexity at which bias and variance are perfectly balanced (Figure 3). To control the model complexity, different regularization parameters have to be set. As it is not possible to calculate the optimal model complexity beforehand, validation methods (as described in section "Model Validation") are used to estimate the best possible parameters for a given dataset. In the following sections, the parameters which were relevant for this study will be briefly discussed.

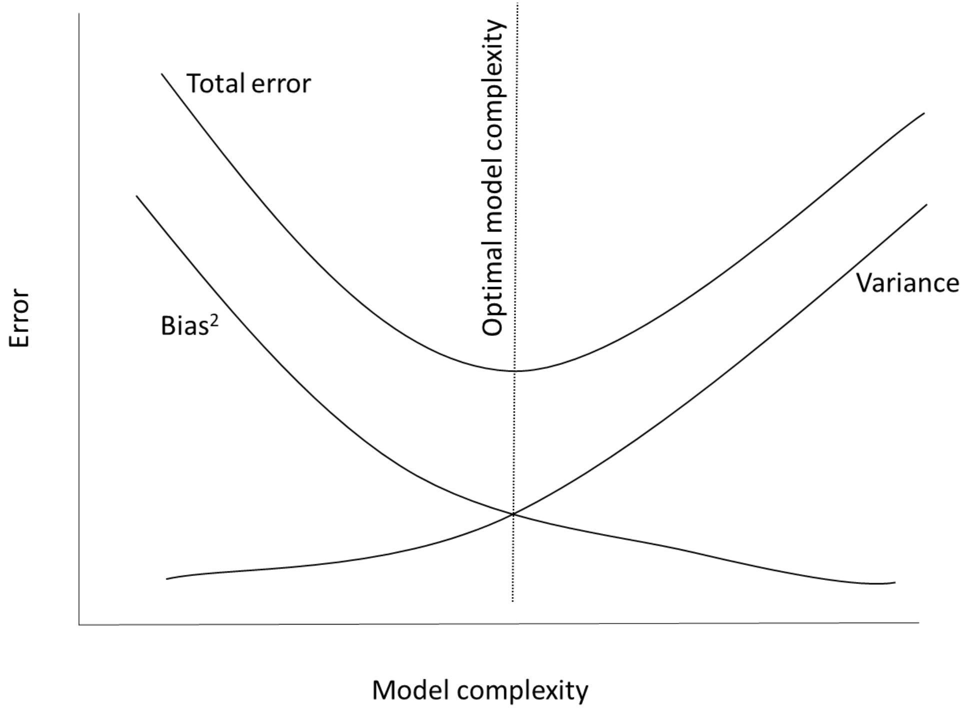

Figure 3. Bias-variance trade-off. Note that the total error is minimized when bias and variance are balanced. At this point optimal model complexity is achieved.

SVM regularization and feature selection

Many regularization approaches exist in the field of machine learning, to control model complexity, avoid overfitting and create generalized models. All regularization methods penalize the weights of features, decreasing them to smaller values. This is done under the premise, that single features with large weights will dominate the prediction and prevent the model from generalizing. Penalizing higher weights therefore results in a simpler model, which ignores noise in the data set. The two standard methods used for SVM are L2-norm and L1-norm regularization. Both approaches are similar but differ in their penalty terms. While L2-norm regularization can reduce feature weights asymptotically close to zero, it cannot eliminate them. L2-norm performs well when all features play a role in the prediction of the label[31]. On the contrary, L1-norm regularization is able to reduce irrelevant features to zero, removing them altogether[31]. This is also referred to as embedded feature selection, because it capitalizes on the weights assigned by the SVM itself. L1-norm works particularly well for data sets with an excess of unimportant features. For the comparison of different atlases, L2-norm regularization was used, because the main interest was to see how different amounts and sizes of parcels would influence the prediction. Therefore we wanted to use a regularization method which would consider all features. For further analysis, L1-norm regularization was applied, because it was assumed that not all features (brain regions), would be predictive for Alzheimer's disease. Additionally this facilitated interpretation of the features and allowed us to investigate selected and non-selected features.

Regularization parameter C

To adjust the strictness of the regularization, one can tune the capacity parameter C. As discussed before, setting the regularization parameters for a model is subject to the Bias-variance trade-off. In our case, the SVM strives to maximize the margin between two classes and avoid any misclassification. However, these two objectives can be contrary. If the given training set contains outliers, SVM will determine a small margin, to avoid any misclassification. Although this will work well for the training data, the resulting model will likely have a high misclassification error for new data, due to the small margin of the SVM. Therefore, Cortes et al. (1995) proposed the idea of a Soft Margin Hyperplane[29]. The regularization parameter C can be used to loosen the misclassification constraints and allow some misclassification, to achieve the best possible fit. A Soft Margin Hyperplane has small values for C, which means that misclassification will not be penalized as strong and a greater margin can be achieved (Figure 4 B). Thereby, although increasing misclassification for the training data, the performance on unseen data will be highly increased. On the contrary, SVM with high values for C will tend to allow no misclassification, by determining a small margin (Figure 4 A). This can result in poor generalizability. Even with linearly separable data sets, smaller values for C can prove beneficial if the sample contains unusual data points.

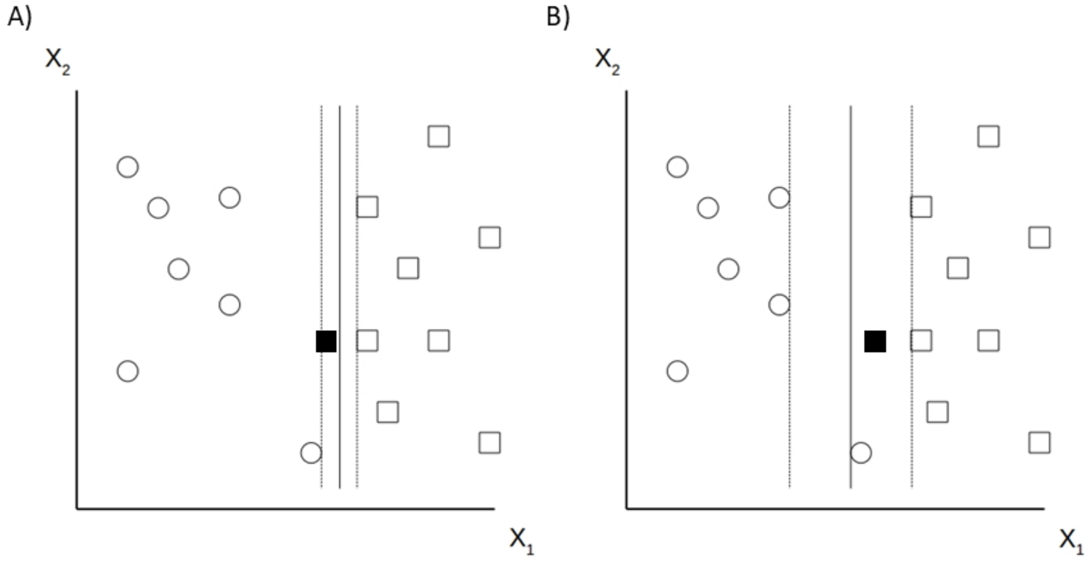

Figure 4. Influence of parameter C. SVM with high C (A) and small C (B). Different symbols (circles and squares) indicate two classes. The filled square represents a new data point. Note that the margin of the classifier is essentially smaller in A than in B. Thereby, the new data point is misclassified in A but not in B.

Hyperparameter optimization

A hyperparameter of a learning model is a variable, which has to be set before training the model on a given dataset. It controls the learning process and influences the model's ability to solve a given task. It is crucial to find the optimal value, which results in the most accurate predictions. More so, tuning the hyperparameter influences the generalizability of the model and is therefore a key point in model building. Since this cannot be done during the actual training, one can implement a preceding cross validation step during which the optimal value is estimated based on the data. The process of finding the best value for a given hyperparameter is called Hyperparameter Optimization[32]. In this study, grid search was implemented to find the optimal value for the regularization parameter C. Grid search is a simple approach, which successively tests predefined values for the hyperparameter to be optimized[32]. This method is also known as exhaustive search, because it tests all possible solutions of a predefined subset, for a given problem. Although simple, grid search comes with a major disadvantage: It quickly becomes computationally expensive with increasing numbers of hyperparameters to optimize. For a number of potential values n, grid search has to solve n optimization problems per hyperparameter. Assuming a fixed n = 10, this means that while tuning one hyperparameter would result in $10^1 = 10$ evaluations, optimizing 4 parameters would multiply to $10^4 = 10{,}000$. This phenomenon is also known as curse of dimensionality, but is avoided in this case, since grid search was only used to find the best possible value for one parameter.

In this study, the values to be tested were defined as 10 evenly spaced values on a logarithmic scale between 10^{-4} and 10^4, in order to cover different magnitudes of positive, as well as negative numbers. The resulting models were then evaluated in a cross validated fashion, where the C-value of the best performing model was passed on to the outer loop, as described in the next subchapter.

Model validation

Validation of the model was based on its ability to generalize. Accuracy, sensitivity and specificity of each model were evaluated on a test set. To estimate the classifiers performance on unseen data, a k-fold cross validation (CV) approach was used. During cross validation, the complete dataset is split into k equally sized subsets. One of the subsets is kept as the validation data, while the remaining k-1 subsets are used for training of the logistic regression model. In successive steps, each subset becomes the validation data once. The advantage of CV is that instead of getting a single evaluation, like when the data is split into one large training and a smaller test-set, one is left with a range of accuracy scores. Therefore, CV is a more reliable estimate of performance. Since the sample was unbalanced, stratified random sampling was used to ensure that both classes are represented proportionally in all folds. When tuning parameters of a model by with CV, one runs into the problem of data leakage, since the test set used for validation is also used to select the value of the parameter in other runs. It is therefore very likely to overestimate the performance of the resulting model. In order to avoid this, nested cross validation was implemented. In the inner cross validation loop, the hyper-parameter C was tested for different values, and the best model was selected. In the outer loop, this model was then evaluated on new randomly created folds.

Balanced accuracy

Datasets in which the two classes are not represented equally are referred to as imbalanced. As in our case, control subjects vastly exceeded the number of AD patients. Therefore, the classifier will often times be very good at predicting the well represented class, in our case healthy controls, but lack accuracy on the underrepresented one. For the conventional accuracy, the sum of correctly predicted samples of both classes is divided by the total amount of subjects. As the model is likely to predict well on the overrepresented class, the conventional accuracy is biased to give overly optimistic estimations in the case of imbalanced datasets. In order to account for this, the balanced accuracy was calculated. The balanced accuracy can be seen as the average accuracy for each class. This means that true positive rate and true negative rate are calculated individually. Subsequently, they are added and averaged. Lower performance on either class will therefore worsen the resulting accuracy, while equally good performance will resolve to the conventional accuracy[33]. Thereby, one gets a more realistic estimate of the models performance.

Permutation scores

To generalize well on new data, it is crucial for a model to detect underlying patterns of the different classes during training. Although the model may predict with good accuracy, it is necessary to validate significance of the obtained results. Therefore, permutation scores were calculated in this study. In order to test if a classification score is significant, classification is repeated on the same dataset, but with randomly permuted labels[34]. The null hypothesis is that features and target variable are independent, and the classifier therefore has not found a significant pattern. The p-value is calculated by the amount of runs which achieved a higher score than the original classification score obtained. If the model has found a significant pattern, accuracies on the permuted labels will be lower, because dependency of features and labels are disrupted. Therefore, the resulting p-value will be highly significant. In this study, permutation scores were calculated in a 10 times 10-fold CV fashion.

Statistical comparison of classifiers over the same dataset

To be able to compare the obtained accuracies of different classifiers, the appropriate statistical test has to be chosen. A paired t-test cannot be applied, since the independence assumption is violated in this case: During cross validation, the same dataset is subsampled to create the various training and test sets, with different sets overlapping with each other. Therefore, a corrected resampled t-test according to Nadeau and Bengio[35] was implemented in this study. By taking the dependency of the different training sets into account, the statistic prevents from underestimating the variance. Thereby, false positive findings are highly reduced.

Evaluating the influence of data representation

Since the primary goal of this study was the evaluation of the influence of data representation, SVM classifiers were trained on different parcellations of the Schaefer atlas. The different atlases were applied to the same pre-processed images of each subject group. Subsequently, one classifier was trained on the features provided by each atlas. Every parcel of each atlas became a feature of the resulting model. Correspondingly, the amount of features ranged from 173 to 1273. This was done once for the 3T sample and once for the 1.5T sample. Models were trained and tested on images with the same magnetic field strength. For this part, L2-norm regularization was chosen to ensure that the different amounts of features would also affect the classification. Otherwise, feature selection could have diminished the effect caused by the different data representations.

Predictive features

To evaluate if SVM can accurately detect disease relevant regions, the regions which were most predictive for the classification of Alzheimer's disease vs healthy controls were analysed. Based on the results of the comparison between the parcellations, the best performing atlas was chosen for further analysis. To facilitate interpretation, feature selection was enforced by implementing L1-norm regularization. The classifier was trained on the exact same data set as before. Subsequently, the weights of the coefficients of the classifier were calculated to detect the regions with the greatest importance for the classification.

Value of non-selected features

To determine the non-selected features, we analysed the weights of all features during each cross validation run. Features which had non-zero weights for at least 80 percent of the folds during CV were deemed selected. Features which had non-zero weights for less than 80 percent of the folds were considered non-selected. To ensure that information contained in the non-selected features would not be redundant with respect to the selected features, the correlation of each non-selected feature with every feature from the selected group was calculated. Pearson's correlation coefficient (PCC) was chosen to determine the correlation[36]. Only features with a correlation lower than 0.3 were included in the non-selected feature group. The threshold of 0.3 was chosen, because correlations lower than 0.3 are considered weak correlations[37]. Therefore, regarding the classification of MCI converters vs MCI non-converters, we compared three different models (Figure 5). First, the original model containing all features. Second, the selected feature model trained only on the features selected in 80% of the CV runs by the original model. Finally, the non-selected feature model which was trained only on the non-selected features. All three models were trained on AD subjects and healthy controls and evaluated on AD and MCI cohorts.

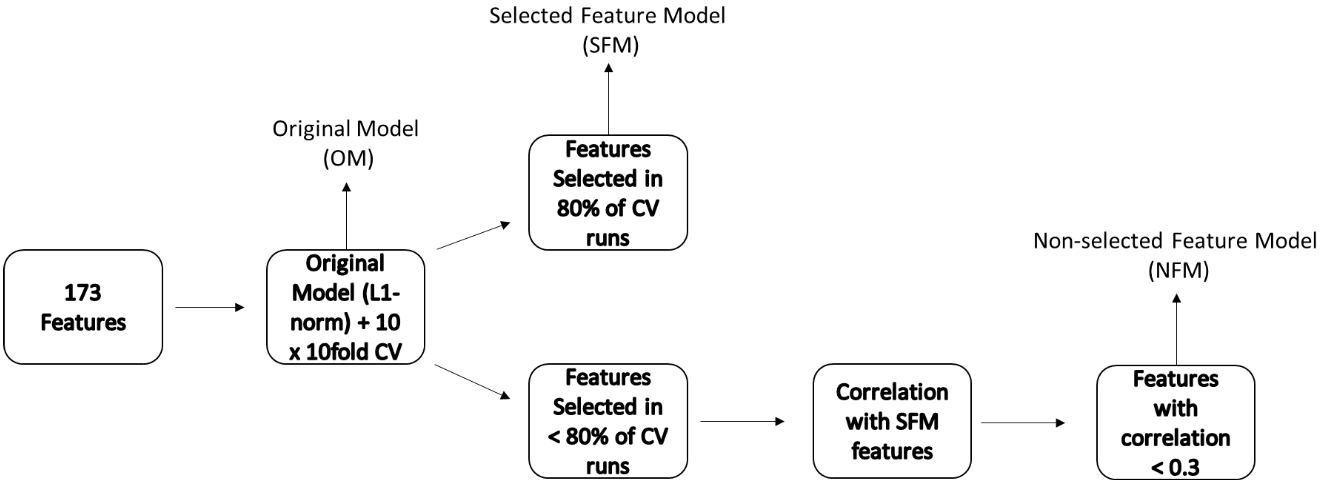
Selected Feature Model
(SFM)
Original Model
(OM)
Features Selected in 80% of CV runs
Non-selected Feature Model
(NFM)
173 Features
Original Model (L1-norm) + 10 x 10fold CV
Features Selected in < 80% of CV runs
Correlation with SFM features
Features with correlation < 0.3

Figure 5. General workflow of building the different models.

Image processing

Processing of the raw images from ADNI was performed by the software program SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12) and the toolbox, CAT12 (http://dbm.neuro.uni-jena.de/cat).

Denoising

The first step of processing magnetic resonance images is to eliminate random noise introduced by the acquisition process. CAT12 handles this by implementing a spatial-adaptive non-local means (SANLM) filter, which estimates local noise and accordingly adjusts the denoising strength of the filter[38]. This has the advantage of removing noise, without also removing high frequency signal components. Another obstacle is presented by the bias field signal, which is a smooth, low frequency signal, caused by magnetic field inhomogeneities within the MRI. These inhomogeneities can result in different intensity values for the same tissue in different parts of the image and therefore impair accurate segmentation. To correct for the bias field signal, a parametric bias correction model is used, which models the different tissue intensities as a mixture of Gaussians[39]. The low frequency portion of the signal is cut off which removes the bias field signal. This is part of SPMs segmentation module, which optimizes signal correction and image segmentation simultaneously.

Spatial normalization and segmentation

In order to correctly identify different tissues in the individual brain scans, SPM implements tissue probability maps (TPM). TPMs are derived from a large set of subjects which are registered to a common space. Since different tissue types are similarly distributed across brains, TPMs represent the probability for a given voxel to belong to a certain type of tissue. This has the advantage that one is not solely reliant on voxel intensities, which can be affected by the partial volume effect if a voxel contains more than one type of tissue. SPM uses tissue probability maps for gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), skull and soft tissue.

To be able to compare identical regions across subjects, all scans need to be brought into the same anatomical space. SPM incorporates spatial normalization and tissue segmentation into the same model: The images are registered to the tissue probability maps and thereby brought into a common space. These maps then represent the prior probability for each class of tissue to be found at a certain region. For segmentation, this information is then combined with the posterior probability, derived from the intensity values of each voxel[40].

CAT-specific preprocessing

CAT12 builds upon this segmentation and further improves it by implementing Local Adaptive Segmentation (LAS), Adaptive Maximum A Posterior estimation (AMAP) and Partial Volume Estimation (PVE). LAS is used to correct for GM-inhomogeneities introduced by varying iron-content or myelinization. While this would typically lead to underestimations of GM at higher intensities caused by high iron content, LAS prevents this by adapting to local intensity differences. In contrast to SPMs segmentation process, AMAP uses the tissue probability maps only for spatial normalization, skull stripping and primary segmentation estimate. Afterwards, AMAP performs segmentation without (further) prior information, by estimating intensity distributions and noise variances based on the local signal intensity[41]. Since basic AMAP does not account for the partial volume effect, CAT complements it by partial volume estimation. Introducing two additional tissue types, GM-WM and GM-CSF, PVE estimates the fraction of different tissue types present in any voxel[42]. This allows for a more accurate segmentation, than tissue classification based on solely the dominant tissue type in a voxel.

DARTEL and Geodesic Shooting

For image registration, DARTEL was used, which stands for “Diffeomorphic Anatomical Registration using Exponentiated Lie algebra”. DARTEL is a large deformation framework which, in contrast to small deformation models, can conserve the topology of the images[43]. The algorithm generates diffeomorphic deformations, which means that the deformations are one-to-one mappings with invertible derivatives. By having invertible derivatives, diffeomorphisms are able to store the relative tissue volume of the brains, reflected in their Jacobian determinants. Thereby, individual differences before and after warping can be preserved. In practice, this means that if a region is stretched during the registration process, the initial signal intensity will be decreased accordingly and vice versa. To make this process more efficient, CAT can implement the Geodesic Shooting approach (GS) [44]. Instead of being reliant on memorizing the entire sequence of velocity fields, GS determines only the initial velocity field and computes deformations on that basis. Therefore, need for disk space is strongly decreased. Additionally, Gauss-Newton optimization is implemented to decrease the number of iterations needed to achieve convergence. GM and WM segments are warped to the DARTEL/Shooting-template in iterative fashion until the registration accuracy cannot be further improved.

Results

Permutation scores

By calculating the permutation scores, all obtained classification scores could be verified as highly significant. Figure 6 shows an example for the calculated permutation scores. Depicted are the scores for the 173 parcel model trained on 3T images of AD vs HC. Since none of the accuracies achieved on permuted labels is close to the obtained accuracy of 91%, the model is deemed highly significant with a p-value of 0.009.

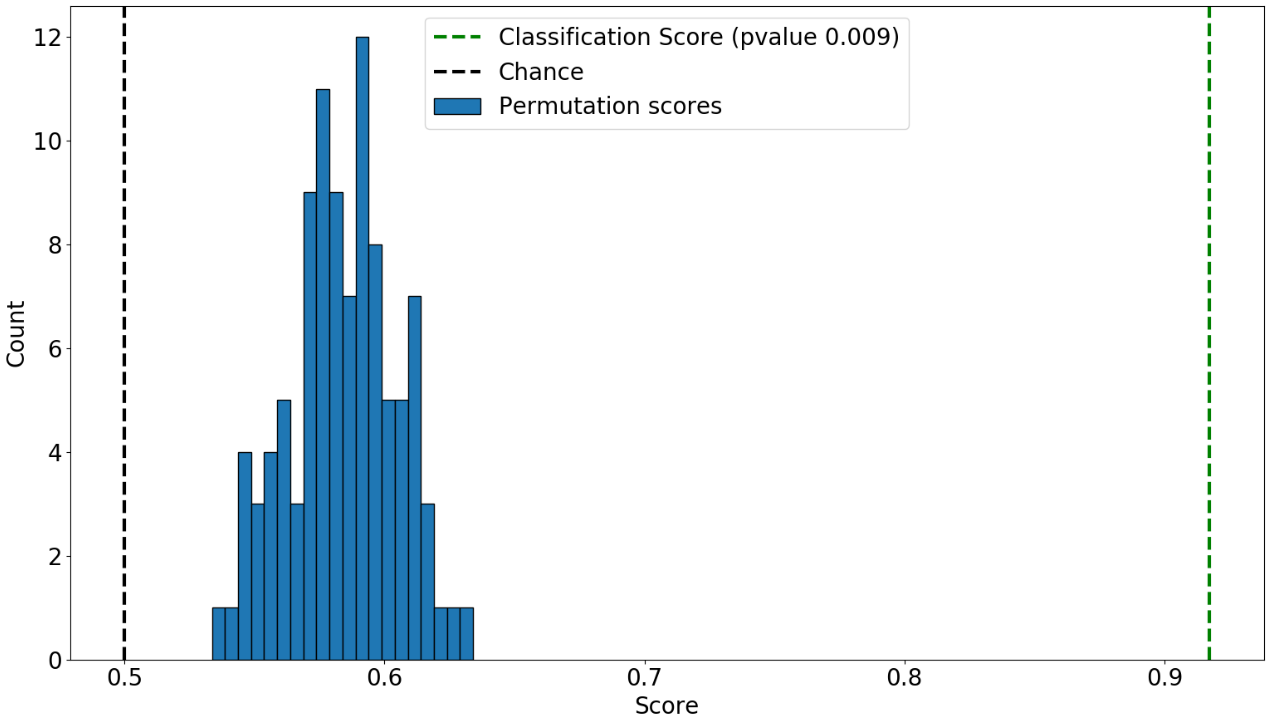

Figure 6. Permutation Scores obtained for the 173 parcel model on 3T images of AD subjects and healthy controls. The black dotted line indicates chance level. The green dotted line indicates the accuracy achieved during the actual classification. The blue bars indicate the amount of runs during which a certain accuracy was observed with permuted labels. Note that none of the bars is close to the achieved accuracy.

Performance of the parcellation-based models

The models were based on nine different brain parcellations. Classification scores, sensitivity and specificity of each model were calculated and are shown in Table 3. For comparison of the models, differences between obtained accuracies over 10 times 10-fold CV were tested for significance. Figure 1 shows boxplots for the accuracies achieved by the models which were trained and tested on 3T images. Although the three models based on the roughest parcellations achieved the highest accuracy, none of the differences between any two models survived p-value correction according to Nadeau. This indicates that classification accuracy of the models did not differ strongly. The best performance was achieved by the 173 parcel model with 91% accuracy, 89% sensitivity and 90% specificity. Lowest accuracy was obtained by the 873 parcel model with 90% accuracy, 88% sensitivity and 92% specificity. All nine models had high sensitivity and specificity and were able to accurately differentiate AD and HC.

On basis of the 1.5 T images, the 173 parcel model performed best as well, achieving an accuracy of 87%, with 84% sensitivity and 90% specificity (table 4). In contrast to the 3T image based models, all models, except for the 473 parcel model, show significantly lower accuracies than the 173 parcel based model. Additionally, the 473 parcel model shows significantly higher accuracies than the 673 parcel model (p-value: 0.04). Apart from that, none of the differences between models achieved significance. In comparison, the sample consisting of only 3T images resulted in superior model performance. Sensitivity as well as specificity are reduced dropping down to 80% and 85%, respectively. Even the worst performing model based on the 3T images (873 parcel model), exceeded the best model out of the 1.5T trained models by 3 percent accuracy.

Table 3
Differentiation rates for SVM-Classification of Alzheimer's disease patients vs healthy controls on 3 Tesla images

Parcellation	Accuracy (%)	Sensitivity (%)	Specificity (%)
173	91.1	88.9	93.3
473	91.1	92.4	89.7
673	90.9	92.4	89.3
773	90.0	90.1	89.9
873	89.8	88.0	91.6
973	90.5	91.0	90.0
1073	90.6	91.3	89.9
1173	90.8	93.4	88.2
1273	90.5	92.4	88.6

Parcellation – indicates the amount of parcels (features) of the atlas which was used to build the classifier

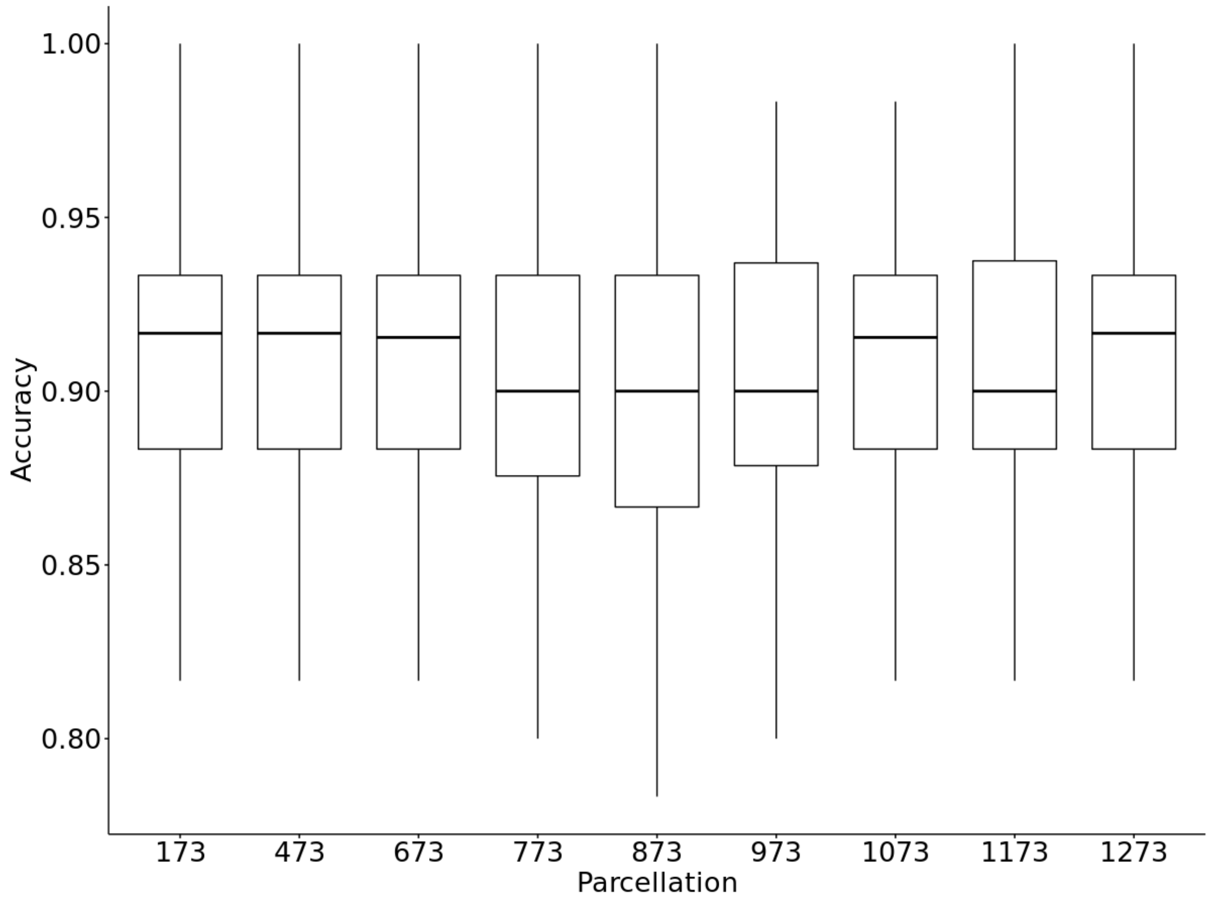

Figure 1 Boxplots of the accuracies achieved by the different parcellation-based models on 3T images

Table 4
Differentiation rates for SVM-Classification of Alzheimer's disease patients vs healthy controls on 1.5 Tesla images

Parcellation	Accuracy (%)	Sensitivity (%)	Specificity (%)
173	86.5	83.6	89.5
473	84.6	82.2	87.0
673	82.9	80.6	85.2
773	83.2	80.3	86.2
873	83.0	80.6	85.4
973	83.2	81.4	85.0
1073	82.9	80.9	84.9
1173	83.0	80.3	85.7
1273	82.5	80.1	84.9

Parcellation – indicates the amount of parcels (features) of the atlas which was used to build the classifier

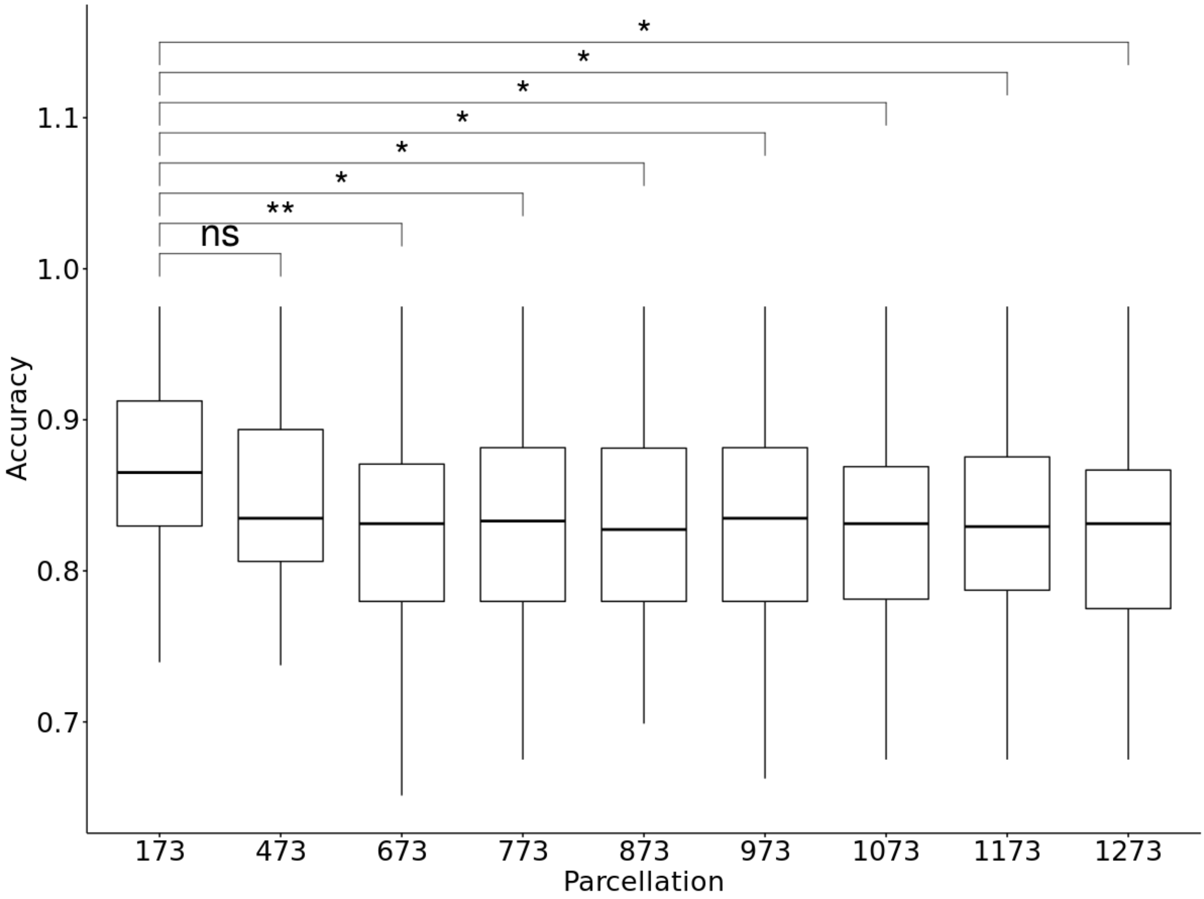

Figure 1 Boxplots of the accuracies achieved by the different parcellation-based models on 1.5T images

Predictive regions

Since the 173 parcel model based on 3T images performed best, it was chosen for further analysis. L1-norm regularization was used to execute feature selection and the model was trained on the 3T sample. The regions which were used for classification by the model are listed according to their importance for the separation of AD and HC and are shown in Table 5. Out of 173 regions, 36 were selected for the differentiation by the L1-norm regularization. The two regions with the highest weights are subcortical regions, more precise parts of Hippocampus and Amygdala.

Table 5
Coordinates of regions used for SVM classification of AD vs HC

Region	Lat	x	y	z	SVM Weight
Hipp_L_2_2	L	-28	-30	-10	-0.33
Amyg_R_2_1	R	28	-3	-20	-0.24
VisCent_ExStr_3	L	-25	-88	20	0.18

LimbicA_TempPole_2	L	-57	-33	-21	-0.16
SalVentAttnA_ParMed_1	L	-11	-34	45	-0.13
Amyg_L_2_1	L	-19	-2	-20	-0.10
SomMotB_Aud_1	L	-53	-23	8	0.09
Cerebellum_Network_15					0.09
SomMotA_4	R	6	-26	70	0.08
DorsAttnA_SPL_1	R	27	-67	51	-0.08
DefaultB_Temp_2	L	-58	-32	-1	-0.08
Tha_L_8_7	L	-12	-22	13	0.08
SomMotB_S2_2	R	57	-4	11	0.08
SalVentAttnA_ParOper_1	L	-59	-38	29	-0.08
SomMotA_1	L	-39	-23	59	0.07
Amyg_L_2_2	L	-27	-4	-20	-0.07
ContB_Temp_1	R	62	-23	-19	-0.06
Amyg_R_2_2	R	28	-3	-20	-0.06
VisCent_ExStr_3	L	-25	-88	20	-0.06
DefaultB_PFCv_2	R	51	28	0	0.05
SalVentAttnB_PFCl_1	R	32	46	29	-0.05
VisPeri_ExStrSup_1	R	13	-86	29	0.04
Tha_R_8_1	R	7	-11	6	0.04
VisCent_ExStr_2	R	22	-96	-5	-0.04
SalVentAttnA_FrMed_1	L	-6	4	62	0.03
SalVentAttnB_PFCl_1	L	-30	44	30	-0.03
Cerebellum_Network_8					0.03
Cerebellum_Network_8					0.02
TempPar_2	R	57	-26	-2	-0.02
SomMotA_1	R	47	-11	48	0.01
DefaultB_Temp_1	L	-55	-4	-20	-0.01
Cerebellum_Network_4					0.01
VisPeri_ExStrInf_1	L	-17	-60	-7	0.01
VisCent_ExStr_3	R	36	-82	16	-0.01
DefaultA_IPL_1	R	55	-51	31	-0.01
SalVentAttnA_Ins_1	L	-41	-1	-7	0.01

Coordinates are in MNI space (L left, R right). The absolute value of the weight (arbitrary units) indicates the importance of the corresponding region for separation between AD and control subjects relative to other regions. Cerebellum networks encompass both hemispheres and are therefore not denoted with coordinates. Abbreviations: Hipp – Hippocampus, Amyg – Amygdala, Tha –Thalamuas, VisCent – Visual A, LimbicA – Limbic A, SalVentAttnA – Salience Ventral Attention A, SalVentAttnB – Salience Ventral Attention B, SomMotA – Somatomotor A, SomMotB – Somatomotor B, DorsAttnA – Dorsal Attention A, DefaultB – Default B, ContB – Control B, VisPeri – Visual B, TempPar – Tempo parietal

Evaluation of non-selected features

Dividing the features as described in the methods resulted in 26 selected features and 147 non-selected features. The 26 selected features were used to build the selected feature model (SFM). Out of 147 non-selected features, 73 were uncorrelated to all of the 26 selected features and used to build the non-selected feature model (NFM). The SFM showed the best performance with 92% accuracy, 89% sensitivity and 96% specificity (Table 6). The original model (OM) performed comparably with an accuracy of 91%, sensitivity of 87% and specificity of 95%. The non-selected feature model (NFM) classifies AD vs HC with an accuracy of 76.4 %, sensitivity of 63% and specificity of 89%. The NFM dropped 15% in accuracy, compared to the OM and the SFM. It is obvious that this is especially caused by a drop in sensitivity, while specificity remains relatively stable. Implementation of the L1-regularization did not improve accuracy in comparison to L2-regularization used during the first analysis. On the other hand, further restriction to the 26 most selected features could increase accuracy about 1%.

Table 6
Differentiation rates for SVM-Classification of early Alzheimer's disease patients vs healthy controls

AD vs HC			
Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Original	91.0	86.9	95.2
Selected	92.1	88.6	95.6
Non-selected	76.4	63.4	88.9

Original – complete parcellation with all 173 parcels, Selected – model including the 26 most selected features, Non-Selected – model including the 73 features which were rarely used and uncorrelated to the 26 most selected features

Classification of MCI-converter vs non-converter

Classification of MCI-Converters vs subjects with early mild cognitive impairment was most accurate using the original model with all 173 features, which achieved an accuracy of 69% (table 7). SFM and NFM achieved 67% and 66% accuracy, respectively. All three models showed high specificity with 88% on average, indicating that they could reliably detect early MCI non-converters. On the other hand, they lacked sensitivity and were only able to correctly identify 47% of the converter subjects, on average.

Out of the three models, none were able to reliably classify MCI-Converters vs subjects with late mild cognitive impairment (table 8). The highest accuracy of 58% was again achieved by the original model, but was far inferior in comparison to classification of eMCI vs MCI-Converter. Finally, all models had relatively high specificity, 75% on average, but low sensitivity with only 37% on average.

Table 7
Differentiation rates for SVM-Classification of converter vs early mild cognitive impairment

eMCI vs Converter			
Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Original	69.0	48.9	89.1
Selected	67.2	48.2	86.2
Non-Selected	66.1	43.1	89.1

Original – complete parcellation with all 173 parcels, Selected – model including the 26 most selected features, Non-Selected – model including the 73 features which were rarely used and uncorrelated to the 26 most selected features

Table 8
Differentiation rates for SVM-Classification of converter vs late mild cognitive impairment

lMCI vs Converter			
Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Original	57.8	40.9	74.6
Selected	55.9	38.0	73.9
Non-Selected	54.8	33.6	76.1

Original – complete parcellation with all 173 parcels, Selected – model including the 26 most selected features, Non-Selected – model including the 73 features which were rarely used and uncorrelated to the 26 most selected features

Discussion

Performance of the parcellation-based models

The present study compared the influence of different types of data representations on the classification performance for Alzheimer's disease patients vs healthy controls. Representatively, SVM classifiers derived from different granularities of the Schaefer atlas were evaluated on their classification accuracy. It is clearly evident, that the granularity of the atlas influences the performance of the resulting model (table 3 and 4). This can be traced back to the smoothing properties of the atlas. When applying an atlas to a given brain image, each parcel will be the average of the underlying voxels of that area. This effect is much stronger than spatial smoothing itself. The usual FWHM (Full Width at Half Maximum) used for smoothing is two times the voxel dimension. Voxels in our study were 1 x 1 x 1 mm, which would result in a FWHM of 2mm. This is on a much smaller scale than the size of a parcel, which can encompass hundreds of voxels. This is also supported by the fact, that smoothing after pre-processing and before applying the atlas had almost no effect on the obtained accuracies (Appendix A and B). Apart from effect size, most of the assumptions made for spatial smoothing are also true for the application of atlases with different granularities. Several studies have shown the varying effect of smoothing with different kernel sizes[45–47]. The same holds true in this case. Starting from the 173 parcellation, the smoothing effect decreases as the individual parcels became smaller with increasing granularity of the atlases. Therefore, the models, tested on images derived from atlases with different smoothing properties, also perform differently. For the classification of AD vs HC, the best performing models were based on the lowest granularity atlases, hence the cases in which the largest smoothing was done. This can be explained by the matched filter theorem, which states that sensitivity for a given effect is highest if the filter matches the extent of that effect[48,49]. Thereby, a matched filter will maximise the signal-to-noise ratio (SNR) in presence of noise. In this study, the observed effect is the atrophy caused by Alzheimer's disease, more so the difference in brain volume between patients and healthy controls. Brain volume loss during AD is strongly increased, with annual atrophy rates of around 3-4%, in comparison to 0.5% in normal aging[50–53]. Between patients and controls, highly affected regions like the hippocampal formation show differences in the order of 1-2 cm^3 [54,55]. From this it becomes evident, that such a pronounced effect is best matched by a filter with high spatial extent. The effect can also be looked at from the brain region perspective. The hippocampus for example is known to be highly atrophied in AD patients. Although atrophy in this region is highly predictive for AD, we do not gain any additional information by splitting it into a huge number of parcels. Rather, we create an excess of features, which (at best) include redundant information. Therefore, model complexity, and more so variance, is increased and the model will naturally perform worse on unseen data.

Classification accuracy between the two field strengths is considerably different. Performance on the 3T images vastly exceeded that for 1.5T images. This may be due to the fact, that a higher magnetic field strength (MFS) increases the SNR[56–59]. In previous studies, an approximate two-fold increase in SNR could be shown from 1.5T to 3T[57]. Therefore, different MFS result in different estimations of regional volume[60]. Additionally, Chow et al. (2015) found that a higher magnetic field strength resulted in improved detection of hippocampal atrophy patterns in AD[55]. In a study by Samper-González et al. (2018) it was also found, that a higher magnetic field strength is connected with increased classification accuracy for AD vs HC[61]. These observations point to the conclusion, that the increase in accuracy is the result of better atrophy detection, due to increased SNR (induced by higher MFS). In conflict with that, several studies report that the effect of magnetic field strength is small compared to the effect size of the disease[62–64]. Nevertheless, all studies acknowledge that there is indeed a difference in estimation of regional volume/cortical thickness. Another finding supports the conclusion that the difference in SNR could cause the decline in accuracy: For the 1.5T images, there is a strong decrease of accuracy with increasing granularity of the atlas. As discussed before, increasing granularity comes with a reduced spatial extend of smoothing. This may affect the 1.5T images even more, because they have a low SNR to begin with. Accordingly, none of the models trained on 3T images showed significantly different accuracy scores. This indicates that additional smoothing by applying the atlas did not improve SNR to a point at which it benefits classification any further.

Predictive regions

By analysing the predictors of our classifier, we wanted to evaluate if this approach is suited to draw biological conclusions about the relevance of certain regions for AD. The regions selected for SVM classification of AD vs HC are generally in line with Alzheimer’s disease specific patterns found by other studies[51,65,66]. The five most important regions of the classifier are discussed representatively.

Notably, 4 of the 5 most predictive regions are located on the left hemisphere (table 5). This is in accordance with several studies, which report that the left hemisphere is more affected in dementia[67,68]. The region which is most useful for the classification is the caudal hippocampus of the left hemisphere (table 5). The hippocampus is known to be one of the most affected regions in Alzheimer's disease, being highly correlated with cognitive decline and overall disease severity[54,69–71]. Although hippocampal atrophy is not specific to the disease as it also occurs in other dementias, like frontotemporal dementia[72], vascular dementia[73] and Parkinson's dementia[74], it is the most established biomarker in AD. Strikingly, only the caudal part of the hippocampus was used by the classifier. As discussed earlier, the implemented L1 norm regularization tends to select only one of several highly correlated predictors. In this case, the caudal hippocampus of the right hemisphere and the rostral hippocampus of both hemispheres were disregarded, although they are known to be affected in AD[74,75]. In Appendix C it can be seen, that using L2-norm regularization, all parts of the hippocampus are assigned with similar high weights. Note that the caudal hippocampus of the left hemisphere is assigned with the highest weight and also selected for classification by the L1-norm regularization, subsequently. The region may be especially predictive, because this part of the hippocampus includes the cornus ammonis (CA) 1-4. It is well known that neuronal loss begins in CA1 initially and, following the Braak stages[76], progressively affects CA2 – CA4 . Atrophy in these parts of the hippocampus is therefore highly predictive for Alzheimer's disease[51,75,77].

The medial amygdala of the right hemisphere is ranked second by our classifier. Especially in the earlier stages of AD, the Amygdala was shown to be highly atrophied[51,78,79]. Reduction in amygdala volume in AD patients compared to healthy controls is estimated to be between 15-25 % [80–82]. Both, the medial amygdala, as well as the lateral parts have been reported to be affected[82,83].

The third most predictive region lies in the extrastriatal region of the visual cortex, in the left hemisphere. Interestingly, several studies have suggested that a subgroup of AD patients, who have visual complaints, suffer from pronounced involvement of the visual association areas[84–86]. Those patients typically show impairment in higher visuoperceptual processing abilities, including visual attention, perceiving structure from motion and face perception[87]. The predominant disturbance of visual abilities is considered by many as its own syndrome, posterior cortical atrophy (PCA), distinct to AD with visual deficits[88–90]. Nevertheless, researchers agree that the underlying neurodegeneration causing PCA is attributable to AD in the majority of cases[90,91]. It is therefore likely, that the first two regions selected by the classifier are particularly involved in the differentiation between typical AD and controls, while the extrastriatal region of the visual cortex is especially helpful for detection of AD cases with involvement of visual areas.

The medial temporal pole of the left hemisphere is the fourth most predictive region of the classifier. It was shown, that AD patients suffer from severe atrophy of the temporal pole, potentially disrupting neural connections to memory related limbic structures[92]Furthermore, Domoto-Reilly et al. (2012) found a high correlation of impaired performance in the Boston Naming Test and cortical thinning of the temporal pole in AD patients[93].

The medial parietal cortex of the left hemisphere was ranked the fifth most important region for classification. It encompasses parts of the posterior cingulate cortex (PCC), which is known to be affected early during AD[94–96]. A study by Seo et al. (2007) demonstrated, that the PCC is subject to cortical thinning in MCI patients[97]. Difference of cortical thickness in the PCC could also be shown for AD patients vs healthy subjects[98].

Evaluation of non-selected features

Testing the non-selected feature model on the classification of AD vs HC, we observe that the model still achieves reasonable accuracy (table 6). Therefore, it can be concluded that the features used for building this model contain valuable information for the differentiation of both classes. Nevertheless, the SVM assigned them with a weight of zero during the majority of original cross validation runs. Conversely, this does not imply that those features, more so the underlying brain regions, have no relevance for the disease. It suggests that other features are more suitable, or equally suitable, for differentiation[23]. The number of selected features is highly influenced by the regularization parameter C. Smaller values of C will decrease the amount of features selected by the L1-norm regularization[99]. During the hyperparameter optimization in our study, an optimal C value of 0.046 was determined. Consequently, our classifier contains rather small amounts of features, disregarding those with lower importance for the classification. Additionally, also the features correlated to the most predictive features are removed in the NFM. Therefore, it becomes evident, that a subset of completely disregarded features still contains relevant information for the classification of AD vs healthy controls. Strikingly, the NFM struggles with correctly identifying AD patients (low sensitivity) but performs well on healthy subjects. A possible explanation is, that the original model selected most disease relevant regions. Therefore, the NFM, which includes only features apart from those selected by the OM, is less suitable to detect AD. In this case, the three most predictive regions of the NFM are located on the superior temporal gyrus / anterior temporal pole (Appendix D). Atrophy, especially in the temporal lobe, is generally increased in AD in comparison to healthy aging. Interestingly, recent studies found out, that atrophy rates in this region are similar to those measured in healthy aging[100,101]. Therefore, increased atrophy in the superior temporal gyrus, although existent, is less sensitive for detection of AD. This supports the thesis that the OM indeed selected the most relevant regions. On the other hand, the NFM was still able to differentiate both classes, although to a lesser degree, because volume differences in this region provide valuable information. The selected feature model exceeds the performance of the OM. Although the original model already performs feature selection by L1-norm regularization, the SFM is further restricted to contain only features, which were selected in the majority of cross validation runs. Thereby, the SFM is an aggregate of the most predictive features. By further decreasing variance, through the reduction of the number of features, the SFM gains in accuracy and outperforms the OM.

Classification of MCI-converter vs non-converter

The NFM shows the lowest performance for the classification of AD vs HC. Still, we wanted to see if the NFM would outperform the OM and the SFM when predicting MCI-Converter vs non-converters. Potentially, one could imagine a region which is highly affected in the earlier stages of the disease, but which's atrophy rate declines in later stages. That region would be highly suitable to differentiate converters and non-converters, but less useful to differentiate established AD and HC. Such a region would have been disregarded by the OM, but could be included in the NFM.

Evaluating the performance of the three classifiers on the classification of MCI-Converter vs early MCI Non-Converter, it is obvious that all are able to differentiate both classes (table 7). The accuracies achieved are in line with those reported in several studies, using structural images from the ADNI database[102,103]. Other studies reported even higher accuracies around 80% [104,105]. Interestingly, all models have high specificity and predict well on early MCI patients who do not convert to AD. On the other hand, sensitivity is low in each of the models, indicating that classification of MCI-converters did not work as accurate. As only the baseline scans of each subject was included, this was an expected result. Although all subjects included in the converter group converted to AD until now, this is not necessarily reflected in the first scan. Since the MCI-Converter group included patients originally diagnosed as early, as well as late MCI, baseline scans of individual subjects may differ in their degree of atrophy. Therefore, converters originally diagnosed with early MCI are prone to be misclassified as early MCI non-converters.

The NFM did not show improved accuracy in comparison to the other models. This may be due to the fact that around 2/3 of the atrophy patterns typical for MCI are shared with those of AD[106]. Since the SFM and OM both contain the most typical regions for Alzheimer's disease, they are expected to accurately predict on MCI patients. On the other hand, performance of the NFM did not differ strongly. This suggests that it contains relevant regions to differentiate converters and non-converters. Interestingly, Misra et al. (2009) found that MCI converters showed increased atrophy in the superior temporal gyrus, among other regions, compared to MCI non-converters[104]. As discussed before, this region is included in the NFM. Although it is not as distinctively for AD vs HC, it might show greater differences between converters and non-converters in earlier stages. This also highlights the fact, that atrophy patterns in MCI are not restricted to regions which are highly atrophied in AD, but can include other regions, as well.

Lastly, we tried to classify MCI-Converters vs late MCI patients. In comparison to the classification of converters vs early MCI, we observe a strong decrease of sensitivity and specificity, leading to a loss in accuracy of around 11% for all models. This was unsurprising, since the MCI-Converter group included mostly patients originally assigned to the late MCI group. Thereby, the differentiation of the converter group vs the late MCI subjects is naturally more difficult. In general, one can assume that brains of late MCI patients are more similar to those of converters than those of early MCI subjects. In addition, with an annual conversion rate from MCI to AD of 7%, many of the late MCI subjects are anticipated to convert in the near future[6].

Conclusion

In conclusion, we could demonstrate that different representations of the brain can influence the classification of AD vs HC. This was mainly traced back to the granularity of the chosen atlas. Although performance on 3T images was unaffected, different parcellations highly influenced accuracies obtained on 1.5T images. Since not all hospitals have access to 3T scanners, AD cohorts are often a combination of 1.5T and 3T images. Therefore, our findings indicate, that future atlas-based studies should pay more attention to the choice of the brain atlas. Selecting the right brain parcellation could highly increase performance of the given approach. Conversely, unsuited brain parcellations will impair classification abilities of the model. In addition, we observed that images with lower MFS generally resulted in decreased performance. Still, in our view, it would make no sense to exclude images with lower MFS from the trainings data, since this would heavily reduce the generality of the model. Although in an optimized setting all images would be available in the best possible resolution, clinical data is not on that level yet.

Analysing the predictors of our classifier, we could demonstrate that all of the top 5 regions selected are known to be related to the Alzheimer's disease. Therefore, we conclude that SVM classifiers are highly suitable to draw conclusions about the importance of the selected regions for a given disease.

We were able to show that non-selected features may still contain relevant information for the investigated disease. This was demonstrated by the example of differentiation between AD and HC. It also emphasizes the fact, that biologically relevant structures may not necessarily be included in the best performing model. We therefore suggest that future studies pay more attention to non-selected features, if they aim to identify all disease relevant regions.

Finally, we demonstrated that SVM classifiers derived from training samples of Alzheimer's disease patients and healthy controls can accurately differentiate MCI-Converters from early MCI non-converters. Although the NFM did not exceed performance of the original model, we could show that features not selected for classification of AD vs HC, may contain valuable information for the classification of converters vs early MCI. The classification of converters vs late MCI non-converters showed less encouraging results, with none of the models achieving accuracy above 58%.

Limitations

Our current study has several limitations: Apart from different parcel sizes and locations, atlases also differ methodologically. Although this was not an issue in this study, since the same atlas was used in different resolutions, the method to divide the brain could also influence performance. Additionally, atlases are derived from different cohorts. Factors like age-range of the used population, sample size and gender distribution determine the representativeness of the atlas. Therefore, in different settings, influence of the chosen atlas can probably not be attributed only to the granularity of an atlas. Furthermore, images acquired at different scanning sites differ in the applied MRI sequence. Since the MRI sequence determines the voxel intensities, varying sequences will result in different values for the same regions. This could also have affected classification. In addition, it is obvious that various learning algorithms deal differently with various amounts of data. Although the same algorithm was used in this study, it might be of interest to analyse the interplay between learning algorithm and atlas in the future.

Suitability of SVM to draw biological conclusions is limited by the fact, that the L1-norm regularization eliminated relevant features with high correlation to others. Therefore, it is certain that not all relevant brain regions are reflected in the SVMs coefficients. In addition, one has to be careful when speaking of relevance of a certain feature. Although selected features are statistically relevant for the classification, this does not guarantee a causal relationship with the disease. It is therefore crucial, to also verify the relationship of selected feature and target in a biological context.

A further limitation of our approach is, that features deemed non-selected still had a low correlation (<0.3) with some of the most predictive features and were indeed assigned with non-zero weights in a small number of CV runs (less than 20%). Therefore one could argue that the regions deemed non-selected in our study could potentially be selected by the classifier. However, given their low predictive value, they would almost certainly be excluded from the final model, even with a moderate regularization. Additionally, our approach is limited to models with good interpretability, such as the linear SVM. More complex models like deep learning methods aggravate the interpretation, because of increasingly complex interrelations between the features of each layer.

Explanatory power of our study is limited by the fact that our MCI-converter group contained mostly patients initially diagnosed with late MCI. One could therefore argue, that classification in our case is mostly based on differentiation of early and late MCI, and not of converter vs non-converter.

Outlook

For future application we would suggest to apply elastic net regularization[99]. The elastic net regularization combines attributes of L1 – and L2-norm penalty to do grouped selection. While the L1-norm penalty term will allow for variable selection, the L2-term penalty will ensure that highly correlated variables are selected together[107]. Such an approach will select groups of relevant variables together and prevent from disregarding important regions. Thereby, the explanatory power of the model would be highly increased. In addition, the models used to differentiate the MCI cohorts showed low sensitivity and were unable to accurately detect converters. To improve sensitivity in future studies, it might be a suitable approach to further divide the MCI-Converter group according to the duration until conversion to AD. Converter with imminent conversion to AD are more likely to show corresponding atrophy patterns in their baseline scans and will be easier to classify. Complementing this, future approaches should focus on classification of subjects who were originally assigned to the same disease stage. Relying solely on early MCI subjects would also have the benefit that resulting models could predict conversion to AD at an earlier time point. Finally, the models used to classify converters vs non-converters were derived from a training sample of AD patients and healthy controls. Although atrophy patterns in MCI are highly overlapping with those of AD, our own results suggest that there are in fact some regions which play a bigger role in MCI. Therefore, using a training set of MCI converters and non-converters could potentially increase performance of the model.

References

1. World Health Organization. Dementia, 2019. (Accessed July 23, 2019, at https://www.who.int/news-room/fact-sheets/detail/dementia).

2. World Health Organization. The top 10 causes of death: World Health Organization, 2018. (Accessed July 23, 2019, at https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death).

3. 2016 Alzheimer's disease facts and figures. Alzheimer's & Dementia 2016;12:459–509. (http://www.sciencedirect.com/science/article/pii/S1552526016000856).

4. Zanetti O, Solerte SB, Cantoni F. LIFE EXPECTANCY IN ALZHEIMER'S DISEASE (AD), Affective, Behavior and Cognitive Disorders in the Elderly. Archives of Gerontology and Geriatrics 2009;49:237–43. (http://www.sciencedirect.com/science/article/pii/S0167494309002350).

5. Gauthier S, Reisberg B, Zaudig M, et al. Mild cognitive impairment. The Lancet 2006;367:1262–70. (http://www.sciencedirect.com/science/article/pii/S0140673606685425).

6. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia--meta-analysis of 41 robust inception cohort studies. Acta Psychiatr Scand 2009;119:252–65.

7. Scheltens P, Fox N, Barkhof F, Carli C de. Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion. Lancet Neurol 2002;1:13–21.

8. Atiya M, Hyman BT, Albert MS, Killiany R. Structural magnetic resonance imaging in established and prodromal Alzheimer disease: a review. Alzheimer Dis Assoc Disord 2003;17:177–95.

9. Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 2010;6:67–77.

10. Eskildsen SF, Coupé P, Fonov VS, Pruessner JC, Collins DL. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. Neurobiol Aging 2015;36 Suppl 1:S23-31.

11. Savva GM, Wharton SB, Ince PG, Forster G, Matthews FE, Brayne C. Age, neuropathology, and dementia. N Engl J Med 2009;360:2302–9.

12. Sunderland T, Wolozin B, Galasko D, et al. Longitudinal stability of CSF tau levels in Alzheimer patients. Biological Psychiatry 1999;46:750–5.

13. Dukart J, Mueller K, Barthel H, Villringer A, Sabri O, Schroeter ML. Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI. Psychiatry Res 2013;212:230–6.

14. Frölich L, Peters O, Lewczuk P, et al. Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia. Alzheimers Res Ther 2017;9:84.

15. Ota K, Oishi N, Ito K, Fukuyama H. A comparison of three brain atlases for MCI prediction. J Neurosci Methods 2014;221:139–50.

16. Salvatore C, Cerasa A, Castiglioni I. MRI Characterizes the Progressive Course of AD and Predicts Conversion to Alzheimer's Dementia 24 Months Before Probable Diagnosis. Front Aging Neurosci 2018;10:135.

17. Chaves R, Ramírez J, Górriz JM, et al. SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. Neurosci Lett 2009;461:293–7.

18. Cuingnet R, Gerardin E, Tessieras J, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 2011;56:766–81.

19. Asim Y, Raza B, Malik AK, Rathore S, Hussain L, Iftikhar MA. A multi-modal, multi-atlas-based approach for Alzheimer detection via machine learning. Int J Imaging Syst Technol 2018;28:113–23.

20. Liu M, Zhang D, Shen D. View-centralized multi-atlas classification for Alzheimer's disease diagnosis. Hum Brain Mapp 2015;36:1847–65.

21. Long Z, Huang J, Li B, et al. A Comparative Atlas-Based Recognition of Mild Cognitive Impairment With Voxel-Based Morphometry. Front Neurosci 2018;12:916.

22. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 2003;3:1157–82.
(http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf).

23. Isabelle Guyon AE. An Introduction to Variable and Feature Selection 2003.

24. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34:939–44.

25. Schaefer A, Kong R, Gordon EM, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb Cortex 2018;28:3095–114.

26. Fan L, Li H, Zhuo J, et al. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cereb Cortex 2016;26:3508–26.

27. Buckner RL, Krienen FM, Castellanos A, Diaz JC, Yeo BTT. The organization of the human cerebellum estimated by intrinsic functional connectivity. J Neurophysiol 2011;106:2322–45.

28. Yeo BTT, Krienen FM, Sepulcre J, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 2011;106:1125–65.

29. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.

30. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine learning practice and the bias-variance trade-off, 2019. (https://arxiv.org/pdf/1812.11118).

31. Koshiba Y, Abe S. Comparison of L1 and L2 support vector machines. In: Proceedings of the International Joint Conference on Neural Networks 2003, Doubletree Hotel-Jantzen Beach, Portland, Oregon, July 20-24, 2003. Piscataway, N.J: IEEE, 2003. p. 2054–9.

32. Feurer M, Hutter F. Hyperparameter Optimization.

33. 20th International Conference on Pattern Recognition (ICPR), 2010, 23 - 26 Aug. 2010, Istanbul, Turkey ; proceedings; International Association for Pattern Recognition; IEEE Computer Society; International Conference on Pattern Recognition; ICPR. Piscataway, NJ: IEEE, 2010.

34. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. Journal of Machine Learning Research 2010;11:1833–63. (http://www.jmlr.org/papers/volume11/ojala10a/ojala10a.pdf).

35. Nadeau C, Bengio Y. Inference for the Generalization Error. Machine Learning 2003;52:239–81. (https://doi.org/10.1023/A:1024068626366).

36. Benesty J, Chen J, Huang Y, Cohen I. Pearson Correlation Coefficient. In: Cohen I, Huang Y, Chen J, Benesty. J, eds. Noise Reduction in Speech Processing. 1. Aufl. s.l.: Springer-Verlag, 2009:1–4. (Springer Topics in Signal Processing, v. 2). ISBN: 978-3-642-00296-0. (https://doi.org/10.1007/978-3-642-00296-0_5).

37. Statistical Power Analysis for the Behavioral Sciences (2nd Edition). [S.l.]: [s.n.], 1988. ISBN: 0805802835.

38. Manjón JV, Coupé P, Martí-Bonmatí L, Collins DL, Robles M. Adaptive non-local means denoising of MR images with spatially varying noise levels. J Magn Reson Imaging 2010;31:192–203.

39. Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005;26:839–51.

40. Ashburner J. Computational anatomy with the SPM software. Magn Reson Imaging 2009;27:1163–74.

41. Rajapakse JC, Giedd JN, Rapoport JL. Statistical approach to segmentation of single-channel cerebral MR images. IEEE Trans Med Imaging 1997;16:176–86.

42. Tohka J, Zijdenbos A, Evans A. Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage 2004;23:84–97. (http://www.sciencedirect.com/science/article/pii/S1053811904002745).

43. Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage 2007;38:95–113. (http://www.sciencedirect.com/science/article/pii/S1053811907005848).

44. Ashburner J, Friston KJ. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. Neuroimage 2011;55:954–67. (http://www.sciencedirect.com/science/article/pii/S1053811910016496).

45. Jones DK, Symms MR, Cercignani M, Howard RJ. The effect of filter size on VBM analyses of DT-MRI data. Neuroimage 2005;26:546–54.

46. Mikl M, Marecek R, Hlustík P, et al. Effects of spatial smoothing on fMRI group inferences. Magn Reson Imaging 2008;26:490–503.

47. Zuo X-N, Xing X-X. Effects of Non-Local Diffusion on Structural MRI Preprocessing and Default Network Mapping: Statistical Comparisons with Isotropic/Anisotropic Diffusion. PLoS One 2011;6.

48. Turin G. An introduction to matched filters. IEEE Trans. Inform. Theory 1960;6:311–29.

49. van Hecke W, Leemans A, Backer S de, Jeurissen B, Parizel PM, Sijbers J. Comparing isotropic and anisotropic smoothing for voxel-based DTI analyses: A simulation study. Hum Brain Mapp 2010;31:98–114.

50. Fox NC, Scahill RI, Crum WR, Rossor MN. Correlation between rates of brain atrophy and cognitive decline in AD. Neurology 1999;52:1687–9.

51. Pini L, Pievani M, Bocchetta M, et al. Brain atrophy in Alzheimer's Disease and aging. Ageing Res Rev 2016;30:25–48.

52. Thompson PM, Hayashi KM, Zubicaray G de, et al. Dynamics of Gray Matter Loss in Alzheimer's Disease. J. Neurosci. 2003;23:994–1005.

53. Barnes J, Bartlett JW, van de Pol LA, et al. A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. Neurobiol Aging 2009;30:1711–23.

54. Frisoni GB, Ganzola R, Canu E, et al. Mapping local hippocampal changes in Alzheimer's disease and normal ageing with MRI at 3 Tesla. Brain 2008;131:3266–76.

55. Chow N, Hwang KS, Hurtz S, et al. Comparing 3T and 1.5T MRI for mapping hippocampal atrophy in the Alzheimer's Disease Neuroimaging Initiative. AJNR Am J Neuroradiol 2015;36:653–60.

56. Phal PM, Usmanov A, Nesbit GM, et al. Qualitative comparison of 3-T and 1.5-T MRI in the evaluation of epilepsy. AJR Am J Roentgenol 2008;191:890–5.

57. Tardif CL, Collins DL, Pike GB. Regional impact of field strength on voxel-based morphometry results. Hum Brain Mapp 2010;31:943–57.

58. Marchewka A, Kherif F, Krueger G, Grabowska A, Frackowiak R, Draganski B. Influence of magnetic field strength and image registration strategy on voxel-based morphometry in a study of Alzheimer's disease. Hum Brain Mapp 2014;35:1865–74.

59. Fushimi Y, Miki Y, Urayama S-I, et al. Gray matter-white matter contrast on spin-echo T1-weighted images at 3 T and 1.5 T: a quantitative comparison study. Eur Radiol 2007;17:2921–5.

60. Pfefferbaum A, Rohlfing T, Rosenbloom MJ, Sullivan EV. Combining atlas-based parcellation of regional brain data acquired across scanners at 1.5 T and 3.0 T field strengths. Neuroimage 2012;60:940–51.

61. Samper-González J, Burgos N, Bottani S, et al. Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. Neuroimage 2018;183:504–21.

62. Ho AJ, Hua X, Lee S, et al. Comparing 3 T and 1.5 T MRI for tracking Alzheimer's disease progression with tensor-based morphometry. Hum Brain Mapp 2010;31:499–514.

63. Dickerson BC, Fenstermacher E, Salat DH, et al. Detection of cortical thickness correlates of cognitive performance: Reliability across MRI scan sessions, scanners, and field strengths. Neuroimage 2008;39:10–8.

64. Stonnington CM, Tan G, Klöppel S, et al. Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease. Neuroimage 2008;39:1180–5.

65. Schroeter ML, Stein T, Maslowski N, Neumann J. Neural Correlates of Alzheimer's Disease and Mild Cognitive Impairment: A Systematic and Quantitative Meta-Analysis involving 1,351 Patients. Neuroimage 2009;47:1196–206.

66. Rabinovici GD, Seeley WW, Kim EJ, et al. Distinct MRI Atrophy Patterns in Autopsy-Proven Alzheimer’s Disease and Frontotemporal Lobar Degeneration. Am J Alzheimers Dis Other Demen 2007;22:474–88.

67. Loewenstein DA, Barker WW, Chang JY, et al. Predominant left hemisphere metabolic dysfunction in dementia. Arch Neurol 1989;46:146–52.

68. Giannakopoulos P, Kövari E, Herrmann FR, Hof PR, Bouras C. Interhemispheric distribution of Alzheimer disease and vascular pathology in brain aging. Stroke 2009;40:983–6.

69. Jack CR, Petersen RC, Xu Y, et al. Rates of Hippocampal Atrophy Correlate with Change in Clinical Status in Aging and AD. Neurology 2000;55:484–9.

70. Gosche KM, Mortimer JA, Smith CD, Markesbery WR, Snowdon DA. Hippocampal volume as an index of Alzheimer neuropathology: findings from the Nun Study. Neurology 2002;58:1476–82.

71. Tabatabaei-Jafari H, Shaw ME, Cherbuin N. Cerebral atrophy in mild cognitive impairment: A systematic review with meta-analysis. Alzheimers Dement (Amst) 2015;1:487–504.

72. van de Pol LA, Hensel A, van der Flier WM, et al. Hippocampal atrophy on MRI in frontotemporal lobar degeneration and Alzheimer's disease. J Neurol Neurosurg Psychiatry 2006;77:439–42.

73. Bastos-Leite AJ, van der Flier WM, van Straaten ECW, Staekenborg SS, Scheltens P, Barkhof F. The contribution of medial temporal lobe atrophy and vascular pathology to cognitive impairment in vascular dementia. Stroke 2007;38:3182–5.

74. Laakso MP, Partanen K, Riekkinen P, et al. Hippocampal volumes in Alzheimer's disease, Parkinson's disease with and without dementia, and in vascular dementia: An MRI study. Neurology 1996;46:678–81.

75. Maruszak A, Thuret S. Why looking at the whole hippocampus is not enough-a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for Alzheimer's disease diagnosis. Front Cell Neurosci 2014;8:95.

76. Braak H, Braak E, Bohl J. Staging of Alzheimer-related cortical destruction. Eur Neurol 1993;33:403–8.

77. Schönheit B, Zarski R, Ohm TG. Spatial and temporal relationships between plaques and tangles in Alzheimer-pathology. Neurobiol Aging 2004;25:697–711.

78. Poulin SP, Dautoff R, Morris JC, Barrett LF, Dickerson BC. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res 2011;194:7–13.

79. Miller MI, Younes L, Ratnanather JT, et al. Amygdala Atrophy in MCI/Alzheimer’s Disease in the BIOCARD cohort based on Diffeomorphic Morphometry. Med Image Comput Comput Assist Interv 2012;2012:155–66.

80. Whitwell JL, Sampson EL, Watt HC, Harvey RJ, Rossor MN, Fox NC. A volumetric magnetic resonance imaging study of the amygdala in frontotemporal lobar degeneration and Alzheimer's disease. Dement Geriatr Cogn Disord 2005;20:238–44.

81. Barnes J, Whitwell JL, Frost C, Josephs KA, Rossor M, Fox NC. Measurements of the amygdala and hippocampus in pathologically confirmed Alzheimer disease and frontotemporal lobar degeneration. Arch Neurol 2006;63:1434–9.

82. Cavedo E, Boccardi M, Ganzola R, et al. Local amygdala structural differences with 3T MRI in patients with Alzheimer disease. Neurology 2011;76:727–33.

83. Qiu A, Fennema-Notestine C, Dale AM, Miller MI. Regional shape abnormalities in mild cognitive impairment and Alzheimer's disease. Neuroimage 2009;45:656–61.

84. Black SE. Focal cortical atrophy syndromes. Brain Cogn 1996;31:188–229.

85. Brewer AA, Barton B. Visual cortex in aging and Alzheimer's disease: changes in visual field maps and population receptive fields. Front Psychol 2014;5:74.

86. Katz B, Rimmer S. Ophthalmologic manifestations of Alzheimer's disease. Surv Ophthalmol 1989;34:31–43.

87. Jackson GR, Owsley C. Visual dysfunction, neurodegenerative diseases, and aging. Neurol Clin 2003;21:709–28.

88. Mendez MF, Ghajarania M, Perryman KM. Posterior cortical atrophy: clinical characteristics and differences compared to Alzheimer's disease. Dement Geriatr Cogn Disord 2002;14:33–40.

89. Benson DF, Davis RJ, Snyder BD. Posterior cortical atrophy. Arch Neurol 1988;45:789–93.

90. Crutch SJ, Lehmann M, Schott JM, Rabinovici GD, Rossor MN, Fox NC. Posterior cortical atrophy. The Lancet Neurology 2012;11:170–8.

91. Nestor P, Caine D, Fryer T, Clarke J, Hodges J. The topography of metabolic deficits in posterior cortical atrophy (the visual variant of Alzheimer's disease) with FDG-PET. J Neurol Neurosurg Psychiatry 2003;74:1521–9.

92. Arnold SE, Hyman BT, van Hoesen GW. Neuropathologic changes of the temporal pole in Alzheimer's disease and Pick's disease. Arch Neurol 1994;51:145–50.

93. Domoto-Reilly K, Sapolsky D, Brickhouse M, Dickerson BC. Naming impairment in Alzheimer's disease is associated with left anterior temporal lobe atrophy. Neuroimage 2012;63:348–55.

94. Zhang H-Y, Wang S-J, Xing J, et al. Detection of PCC functional connectivity characteristics in resting-state fMRI in mild Alzheimer's disease. Behav Brain Res 2009;197:103–8.

95. Scheff SW, Price DA, Ansari MA, et al. Synaptic change in the posterior cingulate gyrus in the progression of Alzheimer's disease. J Alzheimers Dis 2015;43:1073–90.

96. Jacobs HIL, van Boxtel MPJ, Jolles J, Verhey FRJ, Uylings HBM. Parietal cortex matters in Alzheimer's disease: an overview of structural, functional and metabolic findings. Neurosci Biobehav Rev 2012;36:297–309.

97. Seo SW, Im K, Lee J-M, et al. Cortical thickness in single- versus multiple-domain amnestic mild cognitive impairment. Neuroimage 2007;36:289–97.

98. Fennema-Notestine C, Hagler DJ, McEvoy LK, et al. Structural MRI biomarkers for preclinical and mild Alzheimer's disease. Hum Brain Mapp 2009;30:3238–53.

99. Demir-Kavuk O, Kamada M, Akutsu T, Knapp E-W. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinformatics 2011;12:412.

100. McDonald CR, McEvoy LK, Gharapetian L, et al. Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. Neurology 2009;73:457–65.

101. Chan D, Fox NC, Scahill RI, et al. Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. Ann Neurol. 2001;49:433–42.

102. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 2015;104:398–412.

103. Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging 2011;32:2322.e19-27.

104. Misra C, Fan Y, Davatzikos C. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. Neuroimage 2009;44:1415–22.

105. Filipovych R, Davatzikos C. Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI) 2011.

106. Fan Y, Batmanghelich N, Clark CM, Davatzikos C. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 2008;39:1731–43.

107. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B 2005;67:301–20.

Appendix

Appendix A

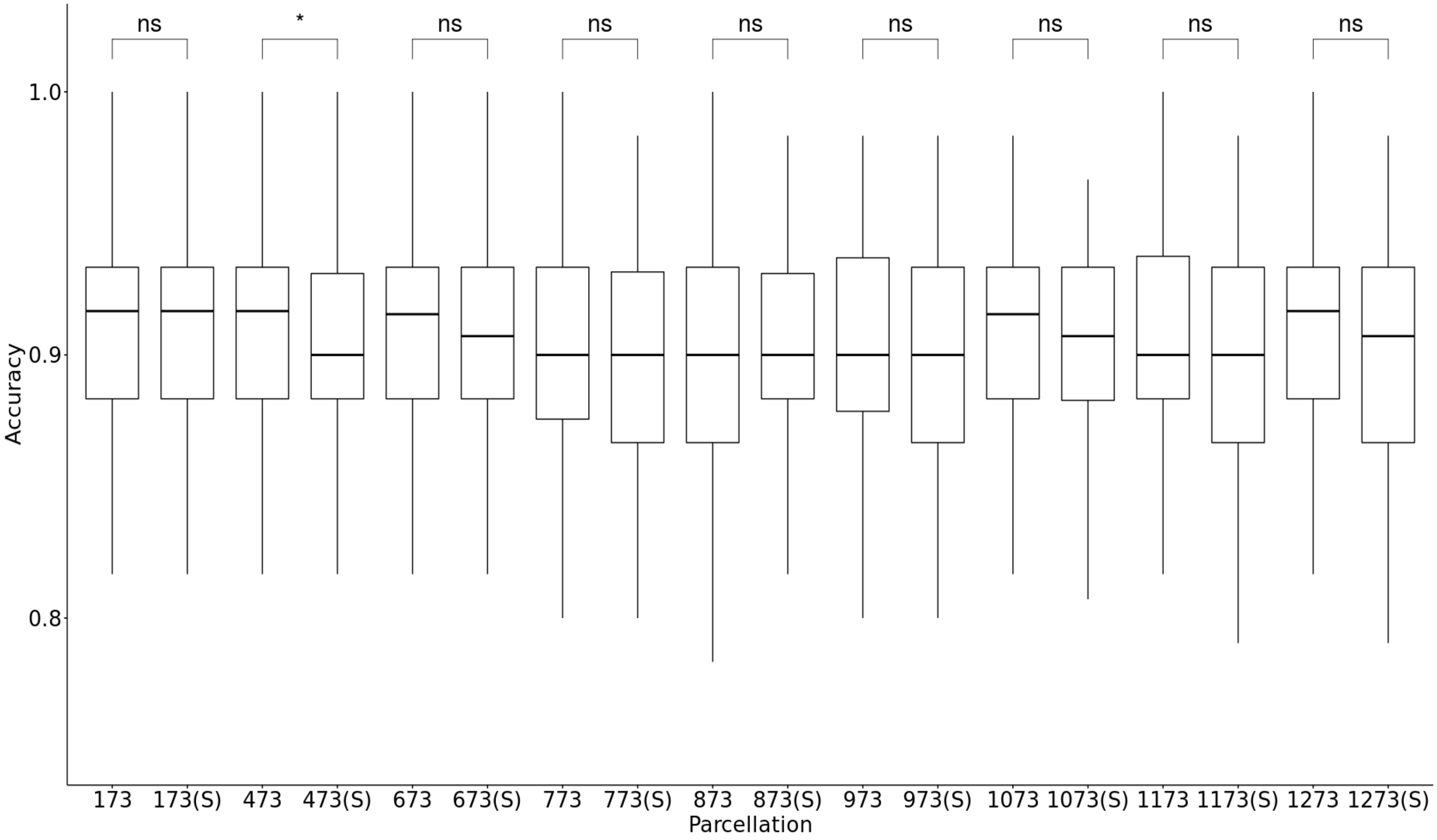

Figure. Comparison of accuracy scores obtained on 3T images with and without smoothing with a 6mm kernel. Accuracies derived from smoothed images are denoted with "(S)". Accuracy scores were obtained by 10 times 10-fold cross validation.

Appendix B

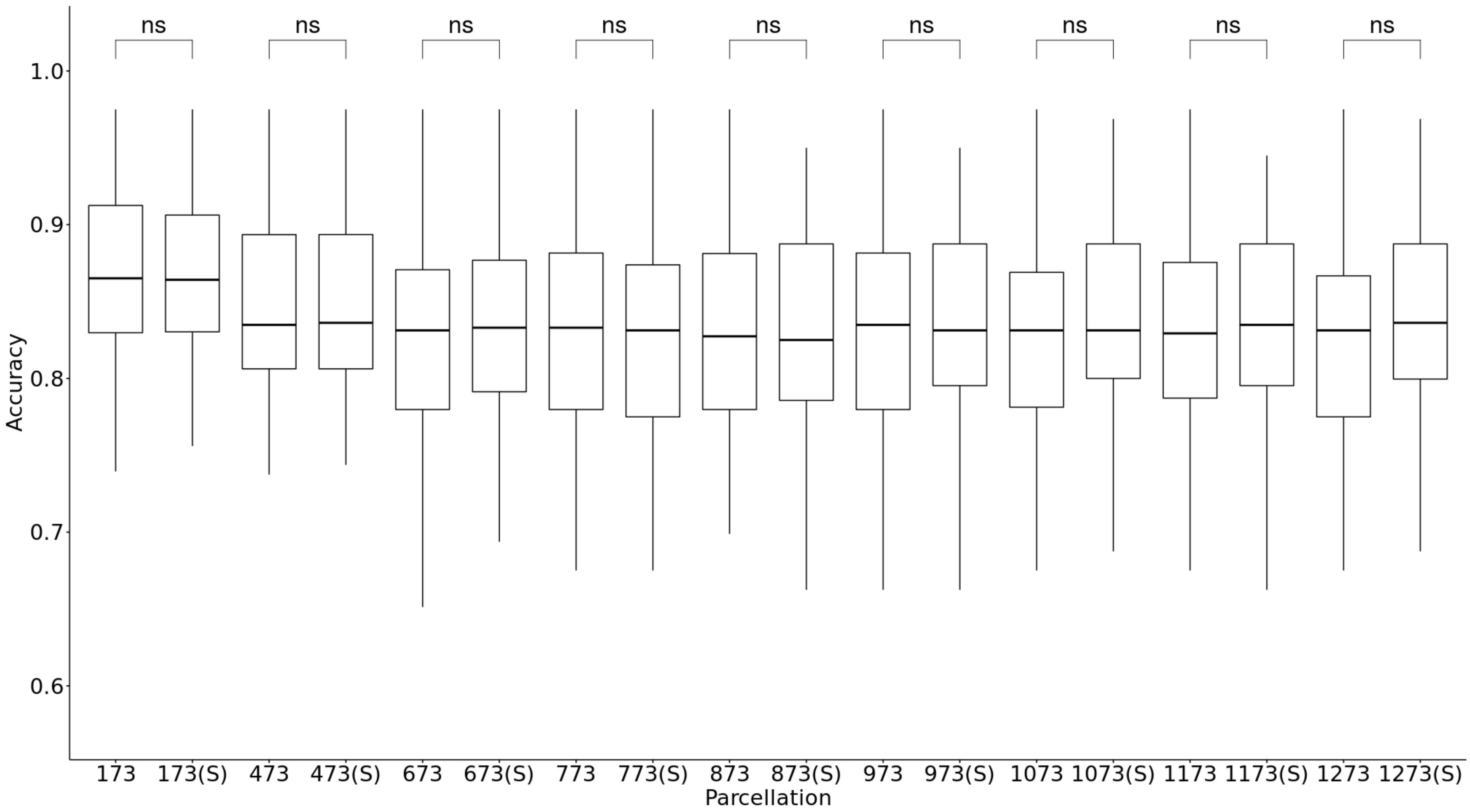

Figure. Comparison of accuracy scores obtained on 1.5T images with and without smoothing with a 6mm kernel. Accuracies derived from smoothed images are denoted with “(S)”. Accuracy scores were obtained by 10 times 10-fold cross validation.

Appendix C

Table 8. Weights assigned to different parts of the hippocampus by the 173 parcel model using L2-norm regularization

Region	Lat	x	y	z	SVM Weight
Hipp_2_2	L	-28	-30	-10	0.15
Hipp_2_2	R	22	-12	20	0.08
Hipp_2_1	L	-22	-14	-19	0.07
Hipp_2_1	R	22	-12	-20	0.05

Coordinates are in MNI space (L left, R right). The absolute value of the weight (arbitrary units) indicates the importance of the corresponding region for separation between AD and control subjects relative to other regions. Hip_2_2 - caudal hippocampus, Hip_2_1 - rostral hippocampus.

Appendix D

Table 9. Coordinates of the top 5 regions used for SVM Classification of AD vs HC by the NFM

Region	Lat	x	y	z	SVM Weight
DefaultB_Temp_1	L	-55	-4	-20	-0.36
TempPar_1	R	51	7	-18	-0.18
TempPar_1	L	-57	-50	12	-0.14
Cerebellum_Network_4					0.09
ventral caudate	L	-12	14	0	-0.08

Coordinates are in MNI space (L left, R right). The absolute value of the weight (arbitrary units) indicates the importance of the corresponding region for separation between AD and control subjects relative to other regions.