TÉCNICO LISBOA

Alzheimer's Disease Modelling through Hidden Markov Models

Diogo Mendes Cardoso

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor: Prof. Maria Margarida Campos da Silveira

Examination Committee

Chairperson: Prof. Paulo Rui Alves Fernandes

Supervisor: Prof. Maria Margarida Campos da Silveira

Member of the Committee: Prof. Ana Luísa Nobre Fred

June 2018

Agradecimentos

Com o término desta dissertação, concluo a minha jornada no Mestrado em Engenharia Biomédica. Para mim, mais do que um curso, são cinco anos recheados de momentos, aprendizagens e amizades que guardarei para o resto da minha vida.

Há percursos que podemos fazer sós, outros que até talvez devamos fazer nessa condição. No entanto, este período é, indelevelmente, marcado pela inestimável presença daqueles que me acompanharam. Por esse motivo, quero, antes de mais, reconhecer e agradecer o seu contributo.

À professora Margarida Silveira, por todas as discussões, sugestões frutíferas e, principalmente, pelo apoio e confiança em mim depositados desde o início. O seu pragmatismo foi determinante para a conclusão desta dissertação.

Aos colegas do Laboratório de Processamento de Sinal do SIPG-ISR, que me acolheram e tiveram sempre uma palavra amiga.

Aos amigos que a faculdade me deu, com quem partilhei, para além desta tese, todo um curso de noitadas a trabalhar, horas a estudar e, principalmente, momentos de amizade e confraternização.

Aos amigos de sempre, com quem há inumeráveis anos tenho vindo a crescer e cuja presença não creio que alguma vez se possa esvair.

Aos meus companheiros do Karaté, que são como uma segunda família, nomeadamente, aos meus mestres que me têm acompanhado quase desde que me lembro.

À minha família, pelo carinhoso sorriso com que me recebe sempre.

Ao meu irmão, por, inconscientemente, me fazer aprender e crescer tanto, todos os dias.

Aos meus pais, por todo o seu percurso de vida. São, verdadeiramente, um exemplo de esforço, exigência, dedicação e perseverança. Sem eles, nada disto teria sido possível.

À Ana, pelo amor...

A todos vós, obrigado por tudo,
Diogo

"Os que passam por nós, não vão sós, não nos deixam sós. Deixam um pouco de si, levam um pouco de nós." – Antoine de Saint-Exupéry

Abstract

In the last few decades there has been a tremendous amount of work in the development of computer aided diagnosis systems (CAD) that can diagnose Alzheimer's disease as early and as accurately as possible, from neuroimaging data such as Magnetic Resonance Images (MRI) or Positron Emission Tomography (PET). The large majority of the proposed methods focused on the diagnosis at a single time instant. Although some authors used longitudinal information, namely, follow-up image data, and reported increased performance, very few works truly explored the temporal evolution. In this thesis we investigate models for disease progression and evaluate their ability to perform diagnosis of Alzheimer's disease. More concretely we developed separate Hidden Markov models (HMM) for modeling the evolution of Alzheimer's disease (AD), Mild Cognitive Impairment (MCI) and cognitively normal (CN) individuals. For each subject, we used PET scans taken at baseline, and at 6, 12 and 24 months follow-ups. We investigate the added value in diagnostic performance of HMM models that capture temporal evolution when compared to diagnosis at baseline.

Keywords: Alzheimer's Disease, Computer-Aided Diagnosis, Positron-Emission Tomography (PET) images, Hidden Markov Models, Disease Modelling.

Resumo

Nas últimas décadas tem havido um enorme desenvolvimento de sistemas de diagnóstico auxiliado por computador que permitam diagnosticar a Doença de Alzheimer o mais precocemente possível com recurso a métodos neuroimagiológicos, tais como Imagem por Ressonância Magnética e Tomografia por Emissão de Positrões (PET). A maioria dos métodos até hoje propostos tem como objectivo o diagnóstico num único instante. Apesar de alguns autores terem mostrado que o desempenho destes sistemas pode beneficiar do uso de informação longitudinal, nomeadamente, a inclusão de imagem médica resultante do acompanhamento regular, poucos trabalhos exploram verdadeiramente a evolução clínica dos pacientes ao longo do tempo. Na presente tese, investigam-se modelos para progressão de doença e avalia-se o seu desempenho no diagnóstico da Doença de Alzheimer. Mais concretamente, desenvolveram-se Modelos de Markov Escondidos (HMM) distintos para a modelação temporal de sujeitos com Doença de Alzheimer, Défice Cognitivo Ligeiro e sujeitos Cognitivos Normais. Para cada sujeito, foram utilizados exames PET adquiridos no inicio do estudo e 6, 12 e 24 meses após, durante o seu acompanhamento clínico. Investiga-se o incremento da capacidade de diagnóstico nos modelos baseados em HMM que incluem a dinâmica temporal em comparação com o diagnóstico com um único exame.

Palavras-chave: Doença de Alzheimer, Diagnóstico Auxiliado por Computador, Modelos de Markov Escondidos, Modelação de Doença.

Contents

List of Abbreviations ix

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Motivation . . . 1

1.2 Proposed Approach . . . 1

1.3 Original Contributions . . . 2

1.4 Thesis Outline . . . 2

2 Alzheimer's Disease: Clinical Background 3

2.1 Introduction . . . 3

2.2 Signs, Symptoms and Pathophysiology . . . 3

2.3 Risk factors, Prevention and Treatments . . . 5

2.4 Auxiliary Diagnostic Tools . . . 6

3 State of the art 7

3.1 Introduction . . . 7

3.2 Modelling temporal evolution . . . 7

4 Proposed Approach 11

4.1 Introduction . . . 11

4.2 Preprocessing . . . 11

4.2.1 Registration . . . 11

4.2.2 Intensity Normalization . . . 12

4.3 Feature Extraction . . . 13

4.4 HMM Learning . . . 13

4.5 Classification . . . 17

5 Experimental Results 19

5.1 Introduction . . . 19

5.2 Dataset 19
5.3 Performance assessment 19
5.4 Classification results 21
5.4.1 2 Classes 21
5.4.2 3 Classes 28

6 Conclusions and Future Work **33**

References **35**

List of Abbreviations

3D Three Dimensional

Aβ Amyloid beta

Acc Accuracy

AD Alzheimer's Disease

ADNI Alzheimer's Disease Neuroimaging Initiative

CAD Computer-Aided Diagnosis

CDR Clinical Dementia Rate

CGM Cerebral Global Mean

CN Cognitively Normal

CSF Cerebralspinal Fluid

EM Expectation-Maximization

GM Gray Matter

HMM Hidden Markov Model

MCI Mild Cognitive Impairement

ML Machine Learning

MMSE Mini Mental State Examination

MRI Magnetic Resonance Imaging

PCA Principal Components Analysis

PET Positron-Emission Tomography

ROI Region of Interest

Sens Sensitivity

Spec Specificity

SPECT Single-photon Emission Computed Tomography

VBSI Ventricular Boundary Shift Integral

VQ Vector Quantization

WMH White Matter Hyperintensities

List of Figures

2.1 Current view on Alzheimer's Disease (AD) onset development. Adapted from McDade and Bateman (2017) 4

4.1 Training procedure for each class model. 11

4.2 Mask used to select the Region of Interest (ROI) for feature extraction. Representation of three cuts in the different anatomical planes. YZ - Sagital, XZ - Frontal and YX - Axial. The remaining picture is a Three Dimensional (3D) ensemble of the represented cuts. . . 14

4.3 Graph representation of a Discrete Markov Model. 15

4.4 Hidden Markov Model (HMM)-based disease model 16

4.5 Classification protocol. 18

5.1 Accuracy obtained with different features as a function of the number of PCA components for each binary classification task. 21

5.2 Sensitivity versus Specificity obtained with different features and number of PCA components for each binary classification task. For the sake of intelligibility only the points with 1,4,7 and 10 principal components are represented. The first point (with one component) is marked with *, and the last (with ten components) is marked with +, the arrows indicate the ascending order of added components. 23

5.3 Accuracy obtained with different features for each binary classification task, considering different time points with a 6-month interval between follow-up scans. 24

5.4 Sensitivity versus Specificity obtained with different features for each binary classification task and considering a 6-month follow-up interval. The result at baseline exam is represented by *, after 6 months by +, and after 12 months by §. 25

5.5 Accuracy obtained with different features for each binary classification task, considering different time points with a 12-month interval between follow-up scans. 26

5.6 Sensitivity versus Specificity obtained with different features for each binary classification task and considering a 12-month follow-up interval. The result at baseline exam is represented by *, after 12 months by +, and after 24 months by §. 27

5.7 Accuracy obtained with different features for the multiclass classification task, considering different time points with a 6-month interval between follow-up scans. 29

5.8 Multiclass Sensitivity versus Specificity, in each of the possible class couplings, for the different feature extraction schemes with a 6-month follow-up interval. The specified class in the subtitle indicates which class was considered separate from the others and whether it was assigned as positive $^{(+)}$ or negative $^{(-)}$. The result at baseline exam is represented by *, after 6 months by +, and after 12 months by §. 30
5.9 Accuracy obtained with different features for the multiclass classification task, considering different time points with a 12-month interval between follow-up scans. 31
5.10 Multiclass Sensitivity versus Specificity, in each of the possible class couplings, for the different feature extraction schemes with a 12-month follow-up interval. The specified class in the subtitle indicates which class was considered separate from the others and whether it was assigned as positive $^{(+)}$ or negative $^{(-)}$. The result at baseline exam is represented by *, after 12 months by +, and after 24 months by §. 32

List of Tables

5.1 Clinical (Mini Mental State Examination (MMSE), Clinical Dementia Rate (CDR)) and demographic information for each group. 20

Chapter 1

Introduction

1.1 Motivation

According to Alzheimer's Association (2017), AD is the most common type of dementia in the elderly, being responsible for 60 % to 80 % of its cases. Additionally, its incidence is expected to increase due to increasing life expectancy.

AD is a progressive and irreversible condition characterized by memory loss along with the decline of other cognitive functions in the areas of reasoning, attention and language. Although there is no cure for AD, early diagnosis is important to rule out other diseases and for the development of treatments that may delay its progression.

The role of brain imaging is of increasing importance in the diagnosis and several imaging biomarkers have been used including two main modalities: structural Magnetic Resonance Images (MRI) and Positron Emitting Tomography (PET). Together with Computer-Aided Diagnosis (CAD) systems, these imaging biomarkers have been shown to provide accurate and early diagnosis.

Despite their success, the large majority of developed CAD systems have not explored temporal evolution on longitudinal data. Therefore, we propose to investigate models for disease progression and evaluate their ability to perform diagnosis of Alzheimer's disease from brain imaging data at different time points.

1.2 Proposed Approach

We propose to build a CAD system for the diagnosis Alzheimer's Disease at different stages that allows the use of follow-up information to complement its decision when needed. For this purpose we propose to develop separate Hidden Markov models (HMM) for modeling the evolution of Alzheimer's disease (AD), Mild Cognitive Impairment (MCI) and cognitively normal (CN) individuals.

1.3 Original Contributions

Although widely used in other areas such as speech processing, statistical modelling of clinical imaging data with HMM was seldom applied. To our knowledge, never with PET and only a few times using the follow-up information. This methodology is promising and allows for simultaneous diagnosis and prognosis.

1.4 Thesis Outline

The present thesis is organized in the following way:

- In Chapter 2, we present a more detailed description of what defines AD and how it can be clinically diagnosed and monitored.
- In Chapter 3, we include a literature review on Machine Learning (ML) methods for modelling of AD with longitudinal data.
- In Chapter 4, we describe the proposed CAD system description to diagnose and model the temporal evolution of AD with PET.
- In Chapter 5, we present a description of the data used in the experiments and the systems' performance is evaluated and discussed.
- Finally, in Chapter 6, we present the thesis conclusions and suggestions for future work.

Chapter 2

Alzheimer's Disease: Clinical Background

2.1 Introduction

After presenting reasons supporting the importance of CAD systems for AD diagnosis in Chapter 1.1, the current section characterizes the disease, and describes its clinical treatments and available auxiliary diagnosis tools.

2.2 Signs, Symptoms and Pathophysiology

Although, during many years, the presence of dementia symptoms was a sine qua non condition for AD diagnosis, nowadays it is not constricted to that requirement. The current definition is much broader, encompassing all the pathological events that culminate in the disease, even if their effect is silent for many years, a graphical representation of the current pathological view on AD can be seen in Figure 2.1. One of the main difficulties in AD diagnosis, and in any dementia type, is that its common first symptoms are often very similar to the ones present in typical ageing (Scheltens et al., 2016). For instance: frequent memory loss, such as forgetting recently acquired information and important appointments; increased difficulty in daily life planing and simple problem solving, with spatial and temporal disorientation episodes; and communication difficulties, such as troubles finding appropriate words to express themselves and keeping track of a conversation. Apathy and mild depression are also common first manifestations. In advanced disease stages, on top of the upper mentioned symptoms development, behavioural changes, poor judgements, motor impairment – such as: speaking, walking and swallowing – are usually followed. The motor function deterioration leaves patients more prone to infections, mostly because difficulty swallowing increases the chance of malnutrition or food entering in respiratory channel, and general motor inactivation weakens the body's immune system. For these reasons, pneumonia and other infectious diseases are common causes of death among AD patients. (Alzheimer's

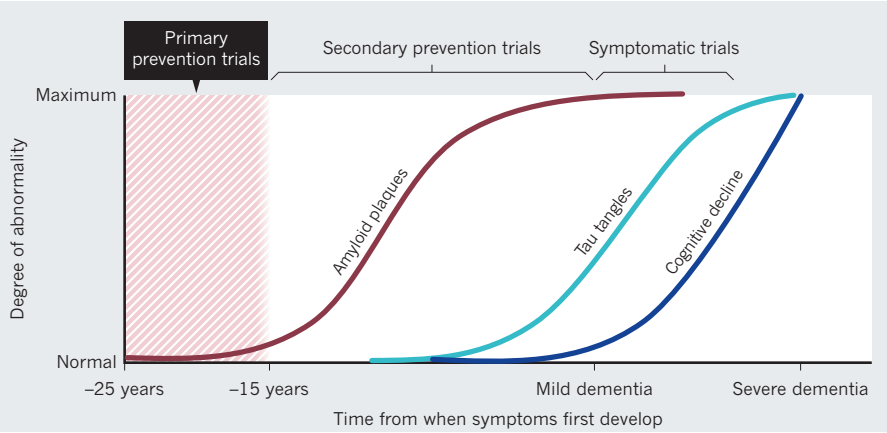


Figure 2.1: Current view on AD onset development. Adapted from McDade and Bateman (2017)

Association, 2017)

The mentioned symptoms result from a series of brain changes under AD. The most characteristic are the cortical accumulation of Amyloid beta (Aβ) in the form of plaques, and the formation of tau protein tangles inside neurons. For a long time, the ruling hypothesis was that these biological compounds were the cause of AD symptoms. Nowadays, although the undeniable relation between these biomarkers and AD, the finding of subjects with Aβ plaques without visible symptoms suggests a more complex relation involving several mechanisms. Nevertheless, Aβ is believed to interfere with synapses, harnessing inter-neuron communication; and tau tangles to affect membrane transport, congesting nutrient and molecule exchange with extracellular medium, thus promoting cellular death and, consequently, brain volume reduction.

Recalling that AD onset is prior to any visible symptoms, as seen in Figure 2.1, and that, after their detection, they have several degrees of severity and ensembles, it is reasonable to define intermediate disease stages. In fact, the acknowledgement of such facts led the scientific community to turn their efforts to a condition known as Mild Cognitive Impairement (MCI), in which a subject has mild but assessable changes in its cognitive abilities, usually perceived only by the themselves or by friends and family. However, this impairment must not affect significantly the subject's ability to perform everyday tasks. Several experts consider that people in this stage are more likely to develop AD or other forms of dementia when compared with Cognitively Normal (CN) ones. In particular, Mattsson et al. (2009) and Visser et al. (2009) made systematic reviews of several studies and concluded that in the time span of 5 years, on average, 32 % to 38% of MCI classified subjects ended up developing AD. The high incidence of AD in this group, provides a reason to give it special attention, stressing the importance of identifying the subjects that will develop the disease. Nonetheless, other forms of the dementia may lead to this stage, and not all subjects reaching this stage will develop any cognitive disease. Moreover, some individuals revert to a CN state or simply stabilize, without progression to AD nor recovery. Its origin may even be iatrogenic[1], as an adverse result from prescribed medical treatments for other conditions.

Despite the MCI clinical cohort relevance, its definition is still too general, leading to a very broad

[1]When a clinical disorder is inadvertently caused by a diagnosis, pharmacological treatment, surgery or any other medical procedure.

and heterogeneous class. This led Alzheimer's Disease Neuroimaging Initiative (ADNI) to establish a more granular scale to classify patients according to their cognitive impairment degree, to get more homogeneous classes and enable a more precise studies with more regular groups. For this, in their second phase, they added two different MCI classes to the previous, an Early-MCI and a Late-MCI, resulting in three classes in between CN and AD classification. In the most recent phase, the Significant Memory Concern class was added as an intermediate stage between CN and Early-MCI.

2.3 Risk factors, Prevention and Treatments

AD has an underlying complex pathological process, which, unfortunateley, is not completly understood. However, there are several factors that are often taken in consideration at diagnosis due to their correlation with the disease's onset probability.

Individuals with at least one AD case in their first-degree relatives have a higher chance of developing the disease, when compared with the ones who do not. And this hypothesis grows stronger with the number of close relatives that manifest the disease, mainly due to possible hereditary and environmental common aspects.

As seen in Section 1.1, AD effects tend to increase in the upcoming years greatly due to the elder population's growth. Thus age is also a risk factor, to be taken in consideration along with family history. The possibility of existing genetic factors implicit when analysing the family's AD cases' existence, has a specific example in a known gene. The APOE gene is involved in a cholesterol-transport protein's transcription. Among the three know alleles of the gene in humans – e2, e3 and e4 – , e3 is the most common, e4 comes right after it and an individual with this form is more likely to develop the disease when compared with the most common, also to develop it at an earlier age; e4 is the rarest, and is known to decrease the probability of AD onset when compared with any of the other two. Considering that each subject inherits one APOE gene from each parent, the combination of the inherited genetic material defines the risk of developing AD.

Beside the already mentioned risks, which, due to their nature, cannot be modified, there is a roll of other risk factors that are more related to everyday life. The prevention of cardiovascular diseases through regular physical activity is one of them. These kind of diseases affect the blood flow and its oxigenation, that is essential for a proper brain function. Mintun et al. (2006) established a relationship between healthy-lifestyle prevention of cardiovascular and dementia diseases.

Ju et al. (2013) concluded that low-quality sleep or its deprivation leads to an increase of Aβ levels in the Cerebralspinal Fluid (CSF). Although not relating them with AD onset, sleep depriveness usually has immediate cognitive effects.

Social and educational aspects have also been proven to be important prevention factors. Individuals with longer formal education and that carry a frequently cognitively challenging life have a reduced risk of developing AD. Besides, an active social life is also beneficial, because demands the brain to actively work in different tasks that are often affected by dementia, such as recalling names and keeping track of a conversation.

Overall, experts believe that it is the upper mentioned factors' combination that determines the risk of developing dementia. Exception being the APOE alleles, that in the cases of e2 and e4 strongly determine how likely it is for a given subject to have (or to develop) AD.

There is no known cure for AD, however, there are several available treatments that ameliorate the symptoms and extent the pre-dementia period. These treatments are either pharmacological, where the drugs are intended to increase the level of certain neurotransmitters, or non-pharmacological where physical exercise and cognitive stimulation take a predominant role.

2.4 Auxiliary Diagnostic Tools

Notwithstanding the importance of typical symptoms' presence, because they are most often the ones triggering pathology's existence awareness, a reliable diagnosis must be based on clinical signs, which are independent from the patient's perception and quantitatively well defined. In order to infer clinical signs, the physicians rely on the screening of some biomarker(s) of interest, i.e. a measurable indicator of some biological state or condition. The use of CDR to measure quantitatively the dementia's symptoms degree, or the MMSE that assesses the patient's cognitive function, although important and widely used, can barely be considered as biomarkers, mainly due to the somewhat subjective principles used in their definition (Waldemar et al., 2007). More definite and reproducible tests are usually based on: the quantification of Aβ and tau protein in the CSF, genetic tests to determine which alleles of the APOE gene are represented in the subjects genome, or on neuroimaging techniques.

Based on the known AD signs, both structural and functional imaging may provide usefull insights in disease assessment. The most common on AD diagnosis are Magnetic Resonance Imaging (MRI) (providing structural information), and Single-photon Emission Computed Tomography (SPECT) and PET (reporting functional patterns).

Nevertheless, the diagnosis is always based on an ensemble of biomarker data, neurological tests, family history and interviews to the patient and relatives, assessing the degree of extent of AD signs, symptoms and risk factors in a patient.

Chapter 3

State of the art

3.1 Introduction

In the last few decades there has been a tremendous amount of work in the development of computer aided diagnosis systems (CAD) that can diagnose Alzheimer's disease as early and as accurately as possible, from neuroimaging data such as Magnetic Resonance Images (MRI) or Positron Emission Tomography (PET). Although the large majority of the proposed methods focused on the diagnosis at a single time instant, there have been several recent attempts to combine cross-sectional and longitudinal information for classification such as Gray et al. (2012), Huang et al. (2017) or Aidos et al. (2017). These methods reported improvements in classification but they didn't truly explore the temporal evolution. On the other hand, several other methods explored the temporal evolution to model disease score progression, without trying to address the question of disease stage classification. This is the case of the works in Chen et al. (2010), Donohue et al. (2014) and Samtani et al. (2012).

In the next section we provide a detailed review of the few methods proposed in the literature for modelling temporal evolution of AD from neuroimaging data.

3.2 Modelling temporal evolution

Sukkar et al. (2012) proposed a method to model AD progression. In this work the main objective was not to classify subjects in a diagnostic group, but rather to uncover more granular disease stages than the ones considered at the time -CN,MCI,AD. It made the assumption that each disease state would correspond to a hidden state, and that the probability of a given observation depends on which state the patient is, this would allow to detect and monitor more subtle changes across time. The emissions consist of four feature vectors that were extracted from ADNI's MRI scans. The first two features were Ventricular Boundary Shift Integral (VBSI) and the Hippocampus volume normalized by the skull volume, the last two were dynamic versions of the previous, i.e., consisted on the change between two consecutive visits of each one (Implying that all the considered individuals in training and testing phase had at least two visits). Assuming that each hidden state accounts for a disease counterpart, a

left-to-right hidden state transition structure is considered. However, this structure is complemented with an allowed right-to-left transition of only one state distance, accounting for possible disease reversions to the previous disease stage. The interval between each scan was 6 months, during a total period of 36 months, although not all the participants had six consecutive scans. The model was assumed to have six hidden states, but no reasoning was provided for that number, it may be to account for six hypothetical transitions between different states in a subject with six scans. The training was done in an unsupervised way, i.e., the subjects' labels were not taken into account so that the model could cluster together similar observations independently of their class. To verify if the clustered in this left-to-right model corresponded effectively to a disease progression one, Viterbi decoding was done for each sequence in order to see in which state each scan was attributed, in the training and in the testing data, and a histogram of the amount of individuals of each class per state was presented. The CN subjects were the most present in the first two classes, decreasing their representation monotonically as the state index rose. The AD were the least represented in the first states, increasing monotonically as state evolved, being the dominant class in the last states. As for the MCI, had their peak in the middle states, being the most represented in them, and decreasing in the peripheral states. This study shows clearly that a HMM AD modelling approach is reasonable, producing interesting results in uncovering more granular disease stages.

Chen and Pham (2013) also presented a promising work using HMM with MRI scans, unlike in the previously explained case, the objective was to distinguish between AD and age related brain changes, also a challenging and important task. The feature extraction methods and model assumptions were also considerably different. There is no real longitudinal information, the temporal series used are fabricated to reflect the Gray Matter (GM) surface structure per slice. The features were extracted using regularity dimension (by Sample entropy) and semi-variograms across slices, combining them in a sequence of slice information accounting for the whole brain, with top to bottom order. The resulting feature vectors' dimension is considered too large for the model's estimation, being so, a Vector Quantization (VQ) method is used to find a smaller set of features to represent the data, using VQ indexes as a low dimensional representation of the vector sequences. A HMM is trained for each class considered - AD, non-demented elder, middle-aged and young-, and, in the testing phase, each individual is assigned to the class in which they have the highest probability to occur. The testing was done in a binary way, in the three possible combinations including AD as one of the classes. Classification between AD and non-demented elder, is the most common type of situation in regular AD and CN comparison, attaining an Accuracy (Acc) of 80.7%, Sensitivity (Sens) of 81.3% and Specificity (Spec) of 80%. The Acc was higher for the other classification problems considered, as expected, having a 92.7% and 98.7% Acc when comparing AD with middle-aged and young classes, respectively.

Wang et al. (2015) took a gait analysis approach were the data was collected using a device including accelerometers and gyroscopes, in order to have movement information along the three possible axis, resulting in a time-series of six-dimensional vectors. An encoding step is responsible for the reduction from six to one dimension observation symbols. Due to the gait natural movement pattern, the HMM follows a cyclic left-to-right topology, with a number of six hidden states, chosen after a state selection

step. The objective is to model gait under single-task and dual-task walking, to study their similarity an detect differences between CN and AD patients. Viterbi decoding returns the most probable hidden sequence originating the observed symbols, after that, is calculated the probability of that sequence to have occurred. This step is done for both of the considered walking tasks and the values obtained are used in a formula that translates the similarity between both. The results show, among others not so relevant for the present work, a significant difference in that score for CN and AD. Providing further evidence that modelling physiological signals and processes through HMM may allow the identification of pathological changes.

Wang et al. (2010) intended to model the volume changes in the brain regions most affected by ageing. For this, they had access to 9-year longitudinal MRI scans of 144 CN subjects. The most relevant regions were determined by an adaptive regional clustering technique based on Pearson's correlation coefficient, selecting the regions whose change would be most correlated with the subjects age. As in most of the analysed models in this context so far, a left-to-right topology was chosen. As for the training, it was done only with CN subjects, because it was meant to model physiological states through ageing. In the testing phase Viterbi decoding was used in order to study the number of transitions, the regularity at which they occurred and the most frequently visited states per subject. At this stage, two groups within the CN were analysed, one where through the follow-up there was no observable decrease in the cognitive function and another where such changes happened. The later, when compared with the first, presented more abrupt state changes, and usually a higher number of transitions in less time, the exception being when the subject was already in the later possible stages. This study shows that while modelling a physiological process it is possible to detect pathological changes when unusual processes within the model happen. This method appears to be a promising tool in brain disease early diagnosis when the patients symptoms are mild and a high degree of uncertainty, in relation to the subject's condition, exists.

Wang and Pham (2011) aimed to make MRI-based age prediction. For motivation includes the possibility of future applications in AD and other neurodegenerative diseases' diagnosis, which, due to cortical thinning and grey matter volume loss (common in ageing), may be addressed under such model. Wavelet-based features were extracted from GM and CSF areas across slices and coded through a VQ method. As in other cases the time-series do not come from longitudinal information, but from the pattern evolution through the brain. A leave-one-out training is performed, with a HMM per subject. For the one to classify, a HMM is obtained in a similar manner. The age attribution is done by choosing, using a Kullback-Leibler divergence based measure, the most similar HMM among the training and it's correspondent age label. In a dataset with 20 subjects, whose ages vary from 50 to 86 years old and the normalized age gap is 2.57 years, the authors claim to accomplish age prediction with an average normalized age prediction error of 2.41 years. By comparing the predicted age and the real age of the patient, one can get an estimate of brain structural degradation and consequently additional information for clinical diagnosis of a pathology.

Pham et al. (2011) starts by explaining that White Matter Hyperintensities (WMH) in MRI are often associated with ageing, AD and other pathologies, and that its quantification is highly dependent on

experts' analysis. The objective is to group individuals with similar WMH patterns, in order to decrease the ambiguity in WMH quantification. The proposed method starts with a segmentation process in order to identify brain tissue and CSF, then areas of WMH, through a thresholding method. The features used are obtained using fractal dimensions on the three previous identified brain areas, followed by a VQ step. This process is repeated for each slice, creating an artificial time-series. A HMM is trained for each individual and, using Kullback-Leibler divergence as a distance measure between models, a similarity matrix is computed, from which, using a phylogenetic tree construction method, a brain tree is attained. With such information, it is possible to cluster subjects according to their scores and detect common characteristics among them. This may be useful in fundamental research, linking WMH values and patterns with certain pathologies (such as AD, whose relation between WMH is still source of debate), and consequently provide further CAD possibilities.

Chapter 4

Proposed Approach

4.1 Introduction

The work developed proposes to investigate the added value in diagnostic performance of HMM models that capture temporal evolution. To accomplish this we implemented a CAD system that models each cognitive stage or class (CN, MCI and AD), as a HMM. Then, given a time-series of brain scans, the system identifies the model under which the sequence is most likely to have been originated from.

Training of each HMM encompasses a number of steps which are depicted in figure 4.1.

Figure 4.1: Training procedure for each class model.

The remainder of this chapter provides a thorough description of all the necessary steps. Section 4.2 describes the preprocessing procedures fundamental for a meaningful image analysis. Section 4.3 encompasses the feature extraction and dimensionality reduction with Principal Components Analysis (PCA). This is followed by a description of the HMM modelling and parameter learning in section 4.4. Finally the classification scheme is presented in section 4.4.

4.2 Preprocessing

Two types of preprocessing were performed, registration and intensity normalization, as described next.

4.2.1 Registration

The imaging data had already been subject to a set of preprocessing procedures, with format, orientation and resolution uniformization purposes. The different scans acquired during a visit had been co-registered to each other and averaged, then the average image was reoriented such that the anterior-posterior axis of the subject was parallel to the AC-PC line and resampled using a 1.5 mm grid. Finally,

the reoriented and resampled image was filtered with a scanner-specific function to produce images with an apparent resolution similar to the lowest resolution scanners used by ADNI.

However, additional preprocessing was needed in order to ensure that images from different subjects are in the same space and voxel-wise comparisons can be performed. Therefore, all the images were warped registered to the MNI standard space (Maldjian et al., 2003), as follows.

First, the brain tissue in all MR images was extracted (skull-stripping) and segmented into white-matter (WM) and gray-matter (GM). The extraction of brain tissue was performed with FreeSurfer (Fischl, 2012). Tissue classification, on the other hand, was conducted with SPM8 (Izquierdo-Garcia et al., 2014) that uses a unified segmentation approach to produce gray and white-matter probability maps.

Second, each PET image was co-registered with the corresponding skull-stripped MR image using SPM8 (Izquierdo-Garcia et al., 2014). Rigid-body transformations (6 degrees of freedom) and an objective function based on the ”sharpness” of the normalized mutual information between the two images were used to conduct these co-registrations. Third, all MR images were non-linearly registered into an inter-subject template using the DARTEL toolbox from SPM8. These templates were then mapped to the MNI-ICBM 152 nonlinear symmetric atlas (version 2009a) using an affine transformation.

Finally, after completing the above steps, the original PET images and the tissue probability maps of GM were resampled into the MNI-152 standard space with a 3x3x3 mm resolution using the appropriate composition of transformations. Morgado et al. (2015) contains a more detailed description of the registration protocol here described.

4.2.2 Intensity Normalization

Intensity normalization is also necessary since it greatly influences the classification process (Dukart et al., 2010; Küntzelmann et al., 2013).

The most common normalization method is the Cerebral Global Mean (CGM), it consists on dividing each voxel's intensity value by the average intensity of all of the intracerebral voxels. However, considering that this average intensity value greatly varies across subjects from different groups (AD patients have a significantly lower average voxel intensity than CN subjects) this normalization process leads to an apparent hyper-activation and hypo-activation in the AD and CN patients images, respectively. Albeit enabling comparison between images, this procedure blurs the differences among the different classes.

From the available protocols, the one chosen was the reference cluster normalization developed by Yakushev et al. (2009), which is a data-driven normalization approach that tackles the attenuation problems in CGM by selecting statistically an intracerebral region that is relatively unaffected across the different classes, i.e. a region preserved during AD progression. From that region, it extracts the intensity reference value to normalize all the images, providing in this way a data-driven normalization procedure that allows image comparison avoiding the inconvenient side-effects expressed in the previous paragraph. This method consists in a first step where CGM normalization is performed. Then, a *t*-test is applied in order to find the hypermetabolic regions in the pathological group when compared with the healthy one. This results in a map where each voxel is defined by a *t*-statistic. A group of voxel

clusters is formed by imposing a threshold on the *t*-value and another on the spatial extent of adjacent voxels. Among the resulting clusters the one with the highest *t*-value voxel is selected as the reference cluster, i.e. the group of voxels that represent the brain areas least affected by AD. The normalization is done relative to the mean intensity value in the reference cluster instead of using all the intracerebral voxels.

4.3 Feature Extraction

In the context of AD, there are specific brain regions where the disease has a greater impact. Therefore, we restricted our image analysis to a set of 10 image regions that were identified by an experienced physician as the most clinically relevant. These are: 1) Left Lateral Temporal, 2) Right Lateral Temporal, 3) Left Mesial Temporal, 4) Right Mesial Temporal, 5) Inferior Frontal Gyrus, 6) Inferior Anterior Cingulate, 7) Left Dorsolateral Parietal, 8) Right Dorsolateral Parietal, 9) Superior Anterior Cingulate, and 10) Posterior Cingulate together with Precuneus. The mask used to select the ROI is depicted in Figure 4.2.

In each of these regions, information related to the voxel-intensity's empirical distribution was extracted, namely average, median and variance.

After feature extraction, PCA (Jolliffe, 1986) is performed independently on the data from each class, representing it in the empirical covariance matrix eigenvectors basis. This stage has two purposes: to decorrelate the data variables (which can have a positive impact in determining the covariance matrices from the hidden states' emissions), and to reduce the data dimensionality – thus decreasing the number of parameters to estimate (that increase quadratically with the number of features). In the developed method, the number of components to retain can be specifically predefined, or automatically chosen following a minimum energy relative amount in the reconstructed vectors.

4.4 HMM Learning

Markov models are stochastic models used to represent random processes and can be used in a wide variety of applications, from meteorology (weather forecast) to bioinformatics (for modelling DNA and protein folding). They rely on the assumption that future states are only influenced by the current ones and not by the events that previously occurred (First order Markov property). These models consider a finite number of states, known as Markov states, and the occurrence of events is represented by transitions between states which have associated transition probabilities. The fact that these transitions are considered time homogeneous and respecting the Markov Property, they can be represented in a transition matrix, describing the transitions between states, and by a vector with the probabilities of each state being the first in a sequence (Rabiner, 1989). A graphical representation of a generic Markov Model can be seen in Figure 4.3.

Hidden Markov Models are methods widely used to model sequences of observations. A HMM is a probabilistic model defined on a discrete finite set of hidden states described by a Markov Model. In

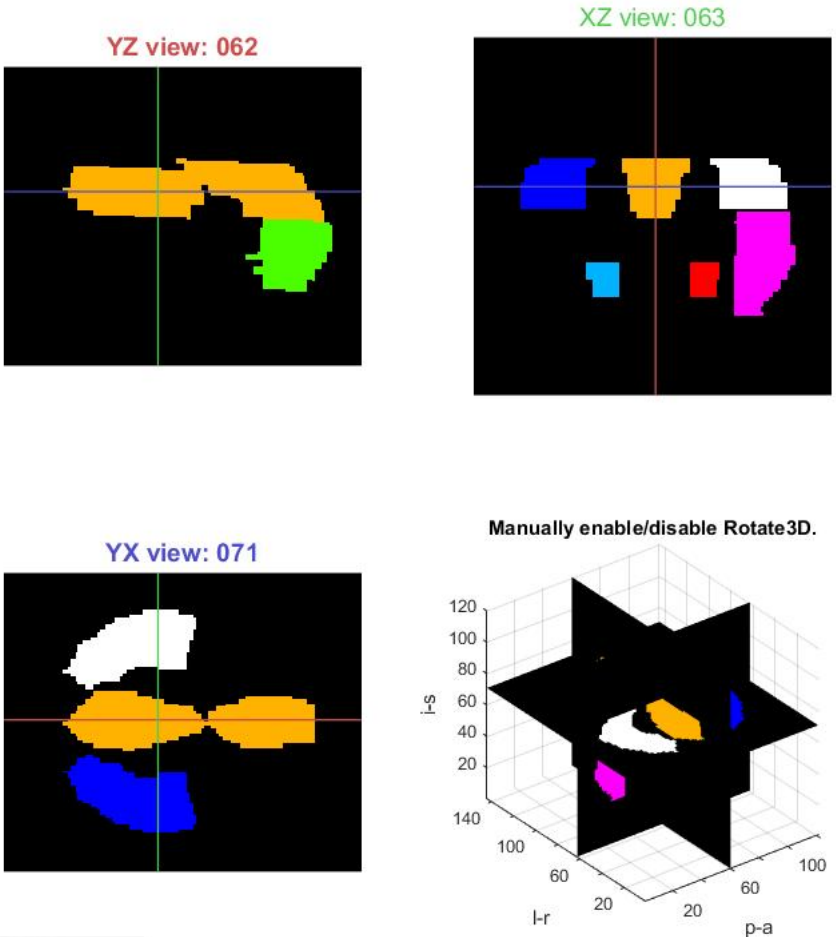


Figure 4.2: Mask used to select the ROI for feature extraction. Representation of three cuts in the different anatomical planes. YZ - Sagital, XZ - Frontal and YX - Axial. The remaining picture is a 3D ensemble of the represented cuts.

HMM, the system's state is not accessible to an outside agent, however, each sample/observation is drawn from a probability distribution that varies according to the system's current hidden state. At every observation, the system's state transition is as described earlier a Markov process. (Rabiner, 1989). For the full description of a HMM, besides the vector defining the initial state probabilities and the transition matrix, it is necessary to know the parameters for each states' emission distribution.

The fact that in different disease stages the patients have altered cognitive behaviours, implies that the accessible set of possible cognitive states is not the same across classes. In accordance with this reasoning, the HMM application in this work focus on modelling each disease stage as a separate HMM. In figure 4.4 there is an illustration of such a model for a specific disease stage. The allowed cognitive states under a certain disease stage are hidden, and we can observe the effect of being in a certain state by observing a PET scan. Given a sequence of PET scans we want to find the model, among the possible ones, in which it is more likely to have been generated.

Considering that it would be unfeasible to the train an HMM that would have as emissions full PET-scans, the model is trained to emit lower dimensional representations of the PET scans based on the

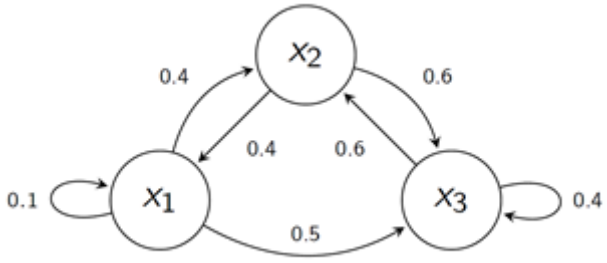


Figure 4.3: Graph representation of a Discrete Markov Model.

feature extraction scheme presented in the previews section.

The feature extraction scheme gives 10-dimensional vectors of continuous values, for this reason the emission type considered for the model's emissions was considered continuous. Among the known continuous multidimensional probability distributions, the Gaussian distribution was chosen mainly due to its simplicity and general use when little is known about the true distribution of the data used (Rasmussen, 2004). A Gaussian distribution is completely defined by its mean vector, and its covariance matrix. In this work the covariance matrix structure is unrestrained, as long as it is positive-definite, to yield a viable Gaussian distribution.

While describing an HMM, it is important to first define the number of hidden states. Afterwards, the probability distributions regarding the initial state and transitions between them, which describe completely the hidden state's dynamics, and the emissions' distribution in each of the hidden states is computed. Although there is an extensive study on methods for the estimation of HMM parameters, such as the one by Lam and Meyer (2010), an alternative protocol was developed in order to create an unsupervised training protocol in order to avoid harnessing the model's ability to describe the analysed temporal processes.

The first part approaches the definition of the number of hidden states and the initial guesses for the parameters of the different distributions to be estimated.

As already stated, in this problem, the emissions where considered continuous following a Gaussian distribution in each hidden state.

There is no reason to assume a specific number of hidden states in each class, or disease stage. Therefore, it is of interest to incorporate their estimation in the learning stage. Since the emission in each state is assumed to be simple, i.e described by a single Gaussian, the collection of several emissions from such model, when analysed all at the same time as individual points (without considering their time relations) may be modelled as a finite mixture of Gaussians. The number of Gaussians in the mixture translates to the number of hidden states in the HMM, and the mixing coefficients result from a combination of the transition matrix and prior probability of each state. Consequently, the problem of determining the number of states is reduced to the one of finding the number of components in a mixture of Gaussians. There are a great number of techniques in the literature that address this problem, the one used in the present work is one proposed by Figueiredo and Jain (2002). This method is an upgrade to the general Expectation-Maximization (EM)-algorithm used to maximize the likelihood

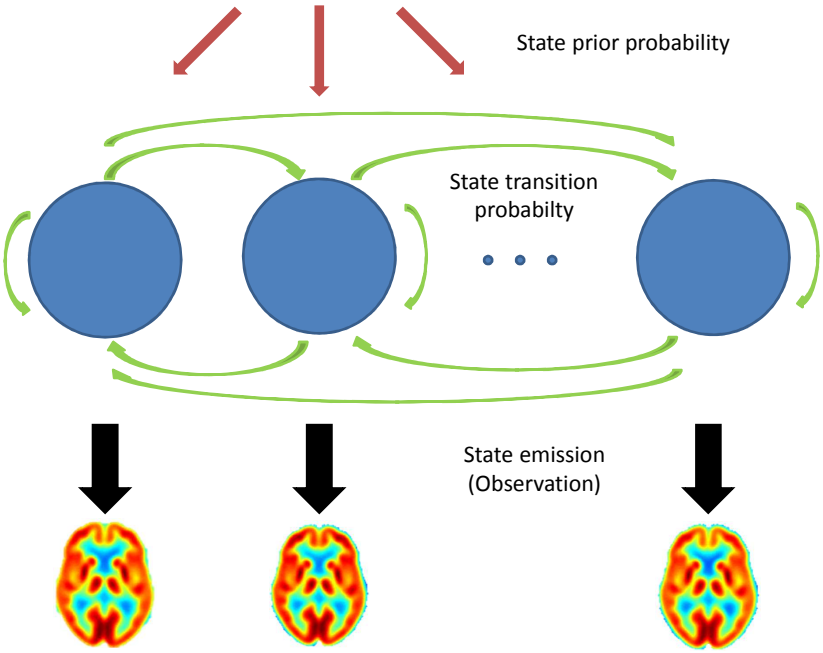


Figure 4.4: HMM-based disease model

function with concepts of information theory, such as the Minimum Description Length, in order to find the appropriate number of components in a finite mixture model and the distribution's best parameters under the estimated number of components. In this step the Gaussian's covariance matrix, as already stated, is considered to be full, i.e. unrestrained.

From here is defined the number of hidden states and the initial mean and covariances guesses for the emissions in each one of them to start the estimation algorithm.

In order to avoid outlier points to harness the training procedure, only the ones yielding a non-zero probability under the mixture model estimated in the previous step are considered.

The initial transition matrix describing the stochastic transitions between hidden states is considered to be uniform and ergodic, i.e. every state has access to every other with uniform probability. This choice was done because imposing a more structured model, such as a left-to-right, would pose the problem of assigning the initial guesses of the Gaussian emission parameters to specific states, and that could deteriorate the model's reliability to describe the class.

The initial guess for the prior probability of starting the process for each state was also considered uniform.

Having defined the number of hidden states and all parameters' starting points, the model's parameters are estimated using the EM algorithm for HMM (Bilmes et al., 1998).

4.5 Classification

After the training phase, there is a vector basis and a set of parameters for each of the three HMM estimated.

In order to classify a sequence of scans, first they have to be written on the basis of each of the classes. Having a model for each class, it is possible to determine the probability of observing a sequence x in class C (equation 4.1).

$$P(x|C) \tag{4.1}$$

Considering $f(x)$ to be the intended classifier's output for the observed sequence x, represented in equation 4.2. However under the model we cannot determine directly such quantities, using Bayes' rule, equation 4.3, taking the HMM as the distribution that describes the observed data, and taking the prior class probabilities $P(C)$ to be uniformly distributed, an approximation of the *Maximum a Posteriori* criterion is used as classifier, depicted in equation 4.4.

$$f(x) \in \arg\max_{C} P(C|x) \tag{4.2}$$

$$f(x) \in \arg\max_{C} \frac{P(x|C)P(C)}{p(x)} \tag{4.3}$$

The value of $p(x)$ does not depend on C, and $P(C)$ is considered uniform to eliminate the effect of class imbalance.

$$f(x) \in \arg\max_{C} P(x|C) \tag{4.4}$$

The scheme represented in Fig. 4.5 summarizes the classification process, given a PET scan sequence of arbitrary temporal length.

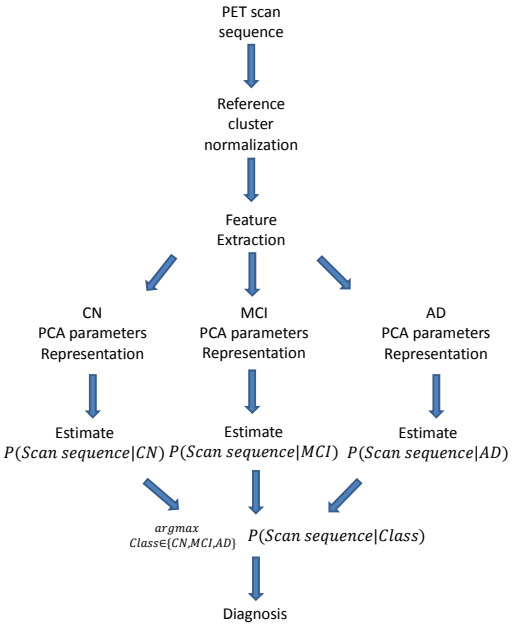


Figure 4.5: Classification protocol.

Chapter 5

Experimental Results

5.1 Introduction

In this chapter, all the experimental details and results from the developed CAD system are presented and discussed. It starts with a description of the dataset used in the experiments, in section 5.2. This is followed by the presentation and discussion of the obtained results, in section 5.3.

5.2 Dataset

The data used in this work was retrieved form the ADNI project's database, in particular from its first study phase - ADNI1 (http://www.loni.ucla.edu/ADNI). The ADNI is a public/private international consortium initiated in 2003, designing, since then, longitudinal studies based on the collection of clinical, genetic, biochemical and neuroimaging data at AD different stages in order to study early detection and disease progression. The dataset consists on PET images of the three clinical groups (CN,MCI and AD), where each patient was monitored during two years, with PET scans taken at baseline, and 6, 12 and 24 months after.

However, not all the patients were monitored in all the time instants previously mentioned. This is mainly due to different follow-up protocols and to the fact that some subjects left the study before completing the period of two years since the first scan at baseline. To allow a better understanding of the dataset used at each time instant, the detailed clinical and demographic information are presented in Table 5.1.

5.3 Performance assessment

In order to calculate the commonly used classification performance measures such as accuracy, sensitivity and specificity, 10-fold cross validation was performed. This procedure is done by dividing the data into 10 folds with similar size. Then, in each iteration, one of the folds is selected to test (testing fold) and

Table 5.1: Clinical (MMSE, CDR) and demographic information for each group.

	CN	MCI	AD
Subjects	75	135	59
Age bas.	75.9 ± 4.6	75.2 ± 7.3	76 ± 6.6
Sex (% of F.)	34.7	35.1	41.4
MMSE bas.	29.1 ± 1.0	27.2 ± 1.6	23.5 ± 2.0
CDR bas.	0.0	0.5 ± 0.1	0.8 ± 0.2
MMSE 6m.	29.1 ± 0.8	26.9 ± 2.4	22.6 ± 3.4
CDR 6m.	0.0 ± 0.2	0.5 ± 0.1	0.9 ± 0.4
MMSE 12m.	29.1 ± 1.2	26.6 ± 2.7	21.0 ± 4.2
CDR 12m.	0.0 ± 0.2	0.5 ± 0.2	1.0 ± 0.5
MMSE 24m.	29.0 ± 1.1	25.8 ± 3.5	19.9 ± 5.1
CDR 24m.	0.1 ± 0.2	0.6 ± 0.3	1.3 ± 0.7

the others are used to train the model (training folds). After all the iterations, the performance measures obtained for each test fold are averaged and the standard deviation is computed.

In this work there are three different classes (CN, MCI and AD) to consider which means that multiclass performance should be computed. However, it is also common in CAD systems to reduce the multiclass problems into sets of binary classification tasks, where all the possible two-by-two class-combination settings are tested. For this reason we present results of binary classification (CN vs AD, CN vs MCI and MCI vs AD) as well as results for the multiclas class problem, which is a task that resembles a set-up closer to the one used in clinical practice.

We present results for different features computed from the voxels-intensity (average, median and variance) in the specified brain regions and for varying number of PCA components on performance.

We also investigate the effect of using different check-up intervals (6 month and 12 month) in the performance of the HMM models. The reason for this separation into two artificial follow-up protocols is due to the fact that the current modelling approach can only deal with uniform time-series, i.e. equally spaced instants, and therefore it is not possible to account, in the same model, for all the exams taken by one patient. With the available data: baseline, 6 months, 12 months and 24 months; it would be necessary to have an exam at 18 months to have an uniform interval of 6 months between exams. Hence the follow-up classification was divided in the two largest possible uniform time-series in the dataset, intervals of 6 months (baseline, 6 months and 12 months) and intervals of 12 months (baseline, 12 months and 24 months).

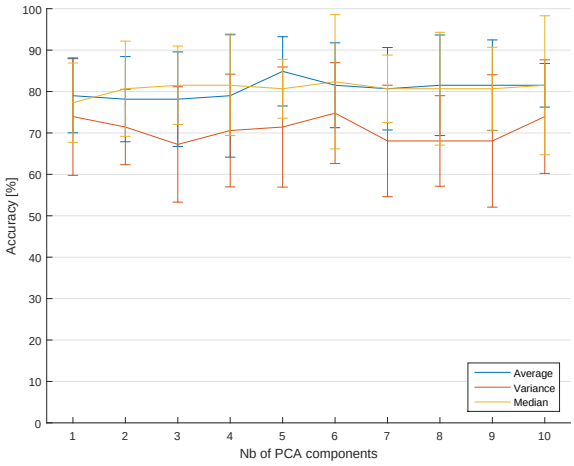
5.4 Classification results

The following two subsections present the classification results obtained for the binary classification tasks as well as for the multiclass task.

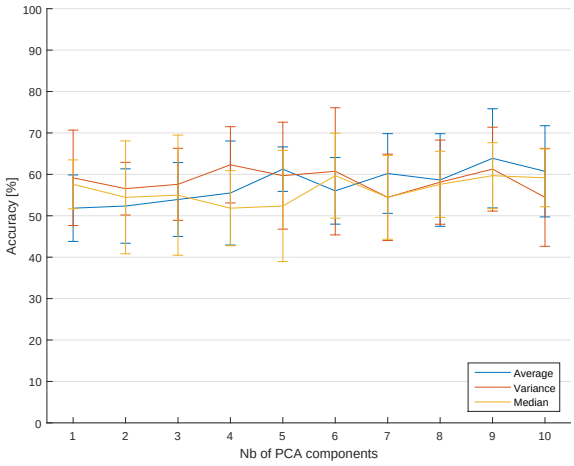
5.4.1 2 Classes

As seen in chapter 4, there is a PCA step in the developed CAD system that acts as a method of decorrelation and dimensionality reduction. The first important parameter to analyse is the number of principal components to use.

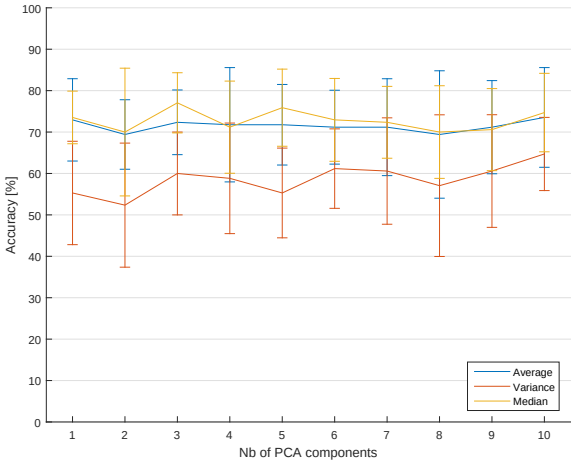
Figure 5.1 shows the classification accuracy (mean and variance) obtained at baseline for varying numbers of principal components used.



(a) CN vs. AD



(b) CN vs. MCI



(c) MCI vs. AD

Figure 5.1: Accuracy obtained with different features as a function of the number of PCA components for each binary classification task.

The overall accuracy is greater in the CN vs. AD task, which is in accordance with the expected results, given that this task opposes the two most dissimilar classes. In general, the variance of intensities

in a region appears to be a considerably worse feature than the average or the median, except in the CN vs.MCI task, where attains similar performance as the other feature extraction schemes. At baseline, the average and the median in the defined ROI demonstrate to be valid features to the addressed problem of automated classification of patients in the possible AD stages. However, neither of them demonstrates to be superior to the other. Considering the Acc line and error bars for the three binary problems, although increasing the number of used principal components in general leads to an increase in Acc, the observed line oscillations prevent any strong conclusion regarding the trend or the ideal number of principal components that maximizes the trade-off between separability and dimensionality. The presence of a subtle peak between 4 and 6 components may be related to the fact that three of the regions are symmetric, meaning that their activation would be more correlated than for totally distinct brain areas. This correlation, in terms of principal components might be interpreted as a direction that describes the overall behaviour of the regions and a term of smaller influence regarding asymmetries.

As for the Sens and Spec analysis, based on fig 5.2, once again, the best overall results concern the CN vs. AD problem. Also in the Sens and Spec domain the variance proves to be the worst feature among the ones studied, although it shows a better Sens both in CN vs. MCI and MCI vs. AD.

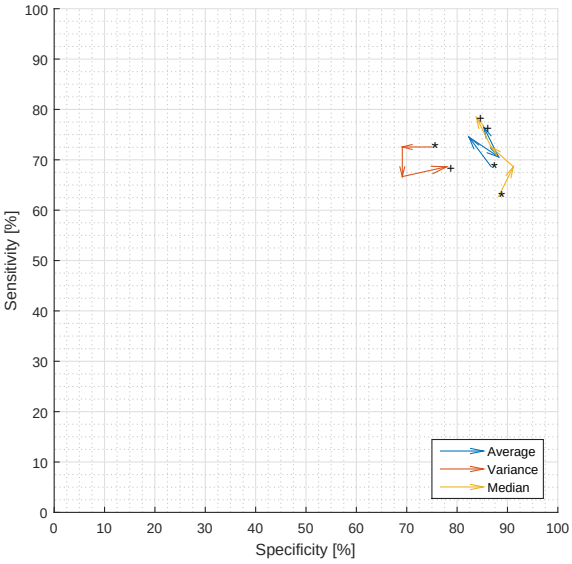
Figure 5.3 presents the accuracy results obtained for the binary classification tasks when different numbers of time points are used, with a 6 month interval between them. The same is shown in figure 5.5 but for 12 month intervals between follow-up scans. The results in both figures were obtained with 10 PCA components.

The plots in Figure 5.3 show that, as expected, accuracy increases with increasing number of time points and that the variability across folds diminishes (smaller error bars). Once again, the CN vs MCI task is the one with the lowest Acc values, and the only where the variance-based features attain similar results to the average and median-based. In one year interval, with two exams, the system reaches an Acc value of 98% for CN vs. MCI, 85% for CN vs MCI, and 93% for MCI vs AD.

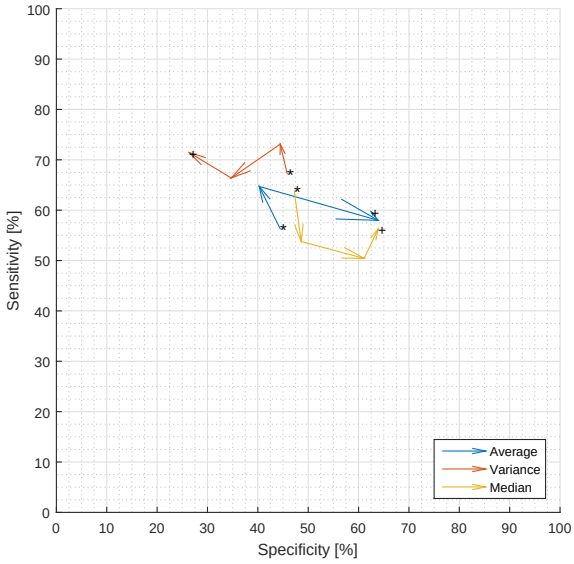
Analysing the Sens and Spec plots in figure 5.4, it is unequivocal that the inclusion of time information improves the systems Sens and Spec, reaching 100% Sens and 95% Spec for CN vs AD, 85% Sens and 86% Spec for CN vs MCI and 85% Sens and 95% Spec for MCI vs AD. The average and median-based features have similar performances, while the variance-based under performs in all the tasks.

The plots in Figure 5.5 and in Figure 5.4 show a similar behaviour with what has been observed in the corresponding plot for 6 month intervals. From the comparison of figures 5.3 and 5.5 we can observe that larger time intervals between exams seems to increase the overall accuracy in all classes, as it would be expected, since the disease tends to evolve and consolidate over time. However, after a one year interval (at 12 months) the results are considerably better with the six month follow-up interval, leading to the conclusion that more important than the time between scans may be the number of scans.

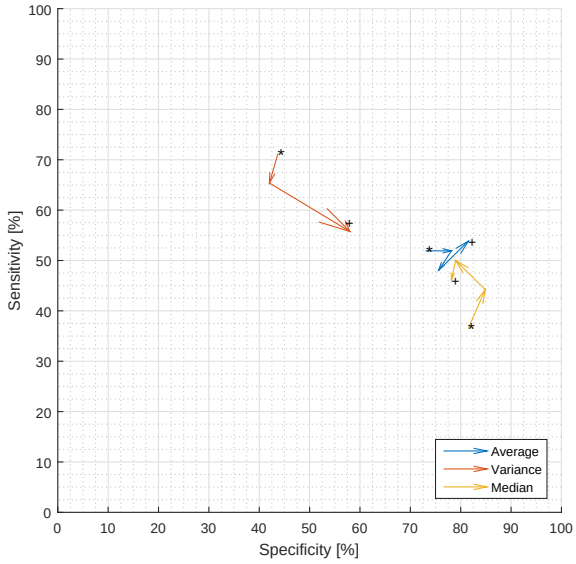
In the CN vs MCI task, the maximum Acc attained for the 12-month interval is lower than the one for the 6-month interval, one reason for this, might be the presence/absence of certain subjects (as explained in 5.2), leading to slightly different datasets, which could make a difference in such a broad and heterogeneous class as MCI.



(a) CN vs. AD

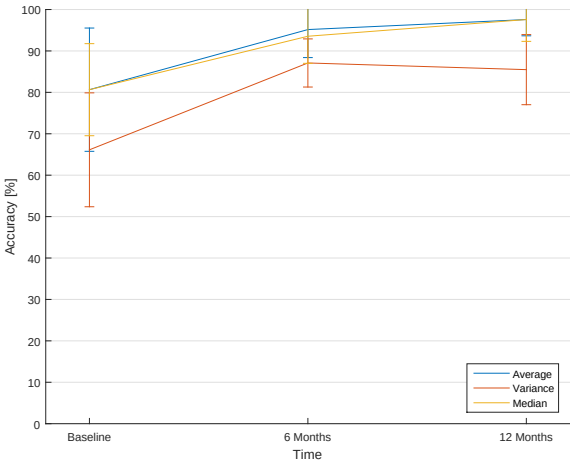


(b) CN vs. MCI

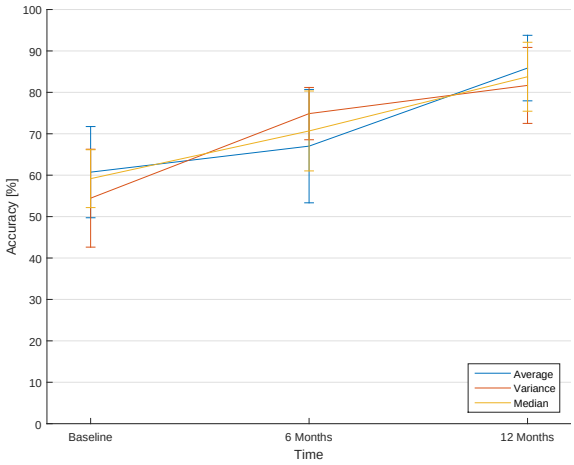


(c) MCI vs. AD

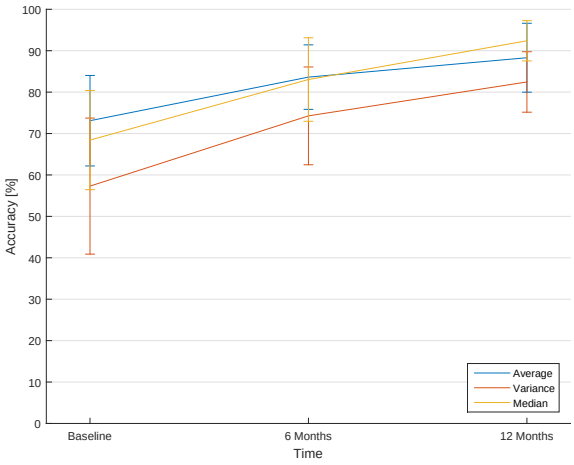
Figure 5.2: Sensitivity versus Specificity obtained with different features and number of PCA components for each binary classification task. For the sake of intelligibility only the points with 1,4,7 and 10 principal components are represented. The first point (with one component) is marked with *, and the last (with ten components) is marked with +, the arrows indicate the ascending order of added components.



(a) CN vs. AD

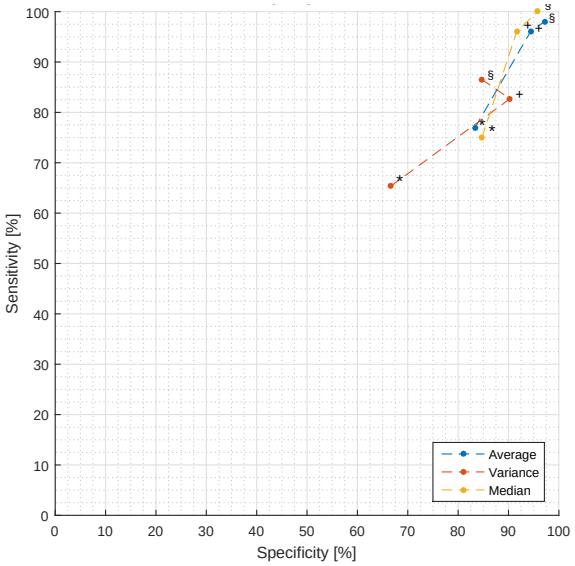


(b) CN vs. MCI

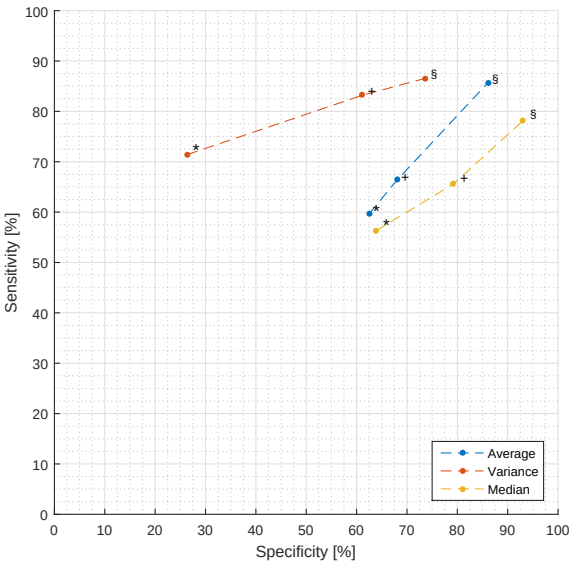


(c) MCI vs. AD

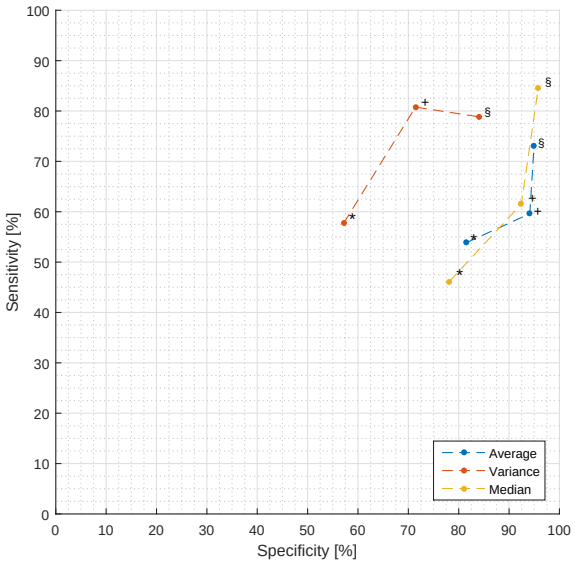
Figure 5.3: Accuracy obtained with different features for each binary classification task, considering different time points with a 6-month interval between follow-up scans.



(a) CN vs. AD

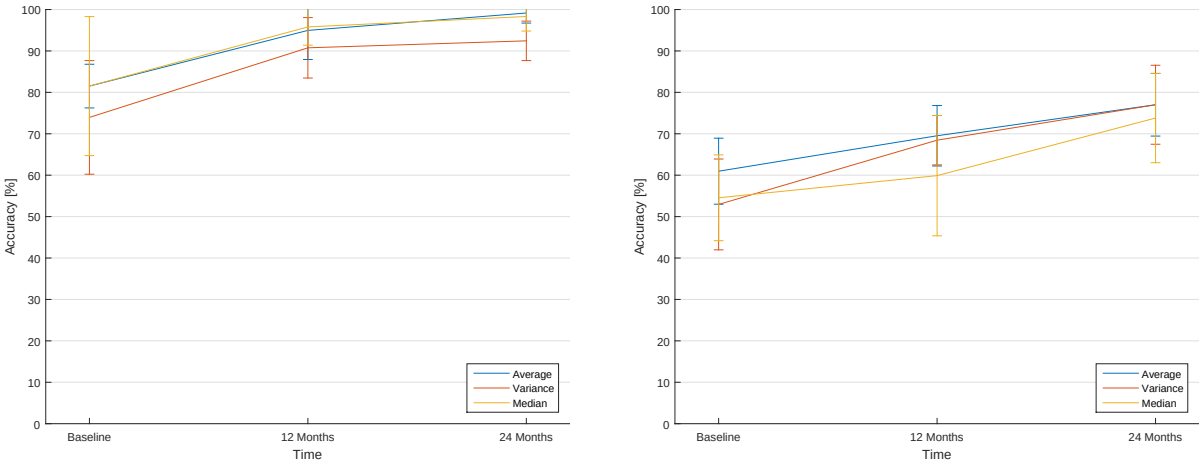


(b) CN vs. MCI



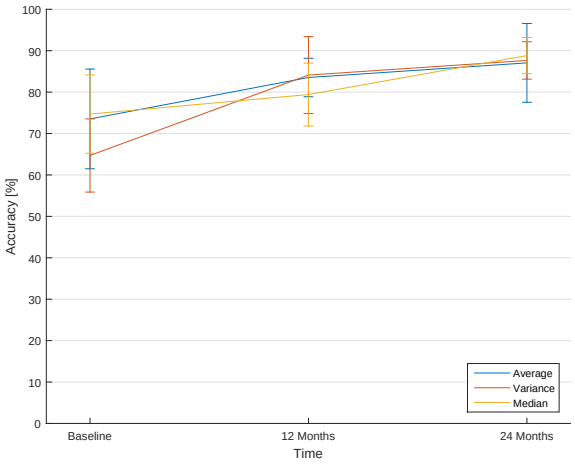
(c) MCI vs. AD

Figure 5.4: Sensitivity versus Specificity obtained with different features for each binary classification task and considering a 6-month follow-up interval. The result at baseline exam is represented by *, after 6 months by +, and after 12 months by §.



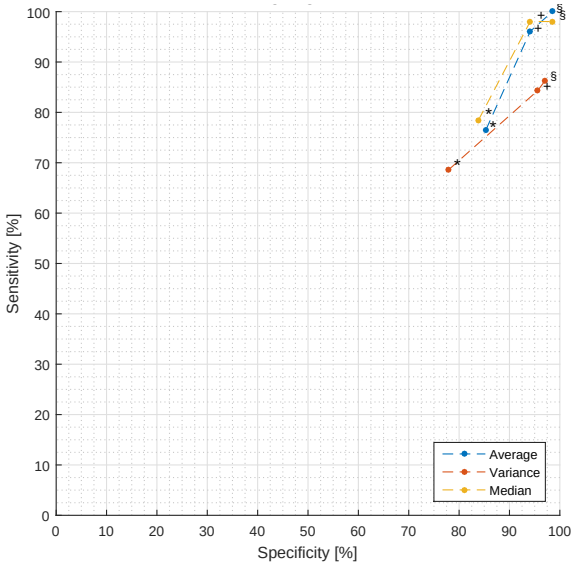
(a) CN vs. AD

(b) CN vs. MCI

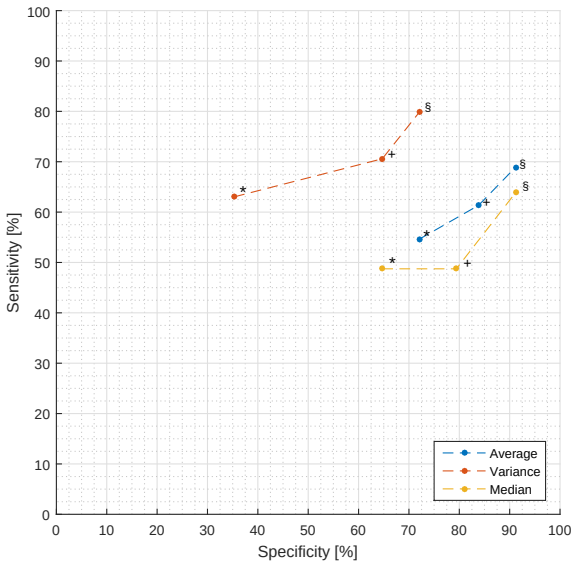


(c) MCI vs. AD

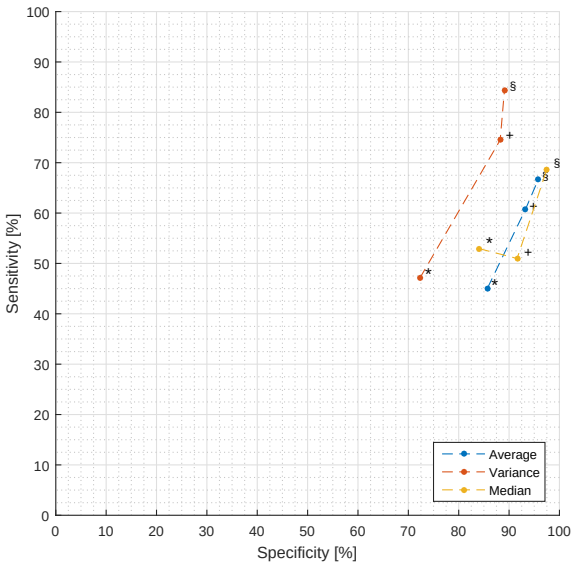
Figure 5.5: Accuracy obtained with different features for each binary classification task, considering different time points with a 12-month interval between follow-up scans.



(a) CN vs. AD



(b) CN vs. MCI



(c) MCI vs. AD

Figure 5.6: Sensitivity versus Specificity obtained with different features for each binary classification task and considering a 12-month follow-up interval. The result at baseline exam is represented by *, after 12 months by +, and after 24 months by §.

5.4.2 3 Classes

Figures 5.7 presents the accuracy results obtained by the multiclass classifier when different numbers of time points are used, with a 6 month interval between them. The same is shown in figure 5.9 but for 12 month intervals.

For the 6-month interval, after a one-year follow up, the system obtained a maximum mean Acc of 63%. As for the 12-month interval, after two years it obtained a maximum mean Acc value of 65%. These values, although considerably lower than any in the binary task are still competitive in AD CAD systems. For the Sens and Spec analysis, the classes were artificially ensembled in order to explore every possible two-by-two groupings, in order to respect these performance assessment parameters definition.

As in the binary task, the groupings took into account the clinical definition of Sens and Spec, were the positive class is the diseased, or in this cases the ones that represent the more advanced disease stage. The three groupings are the following: 1) CN as the negative and MCI together with AD as the positive, 2) AD as the positive and CN together with MCI as the negative, and finally, 3) MCI as the positive with CN and AD as the negative. This last one does not respect the positive/negative stated attribution principle because it is not possible under such grouping, although it should be included in the system's performance analysis.

When considering the CN $^{(-)}$, in the 6-month interval setting, the system has as its best performance values of 85% for Sens and 95% for Spec. Although, with variance-based features the value of Sens is 5% higher than with the other features. Under this setting the system has a high degree of confidence when it classifies a subject in a class other than CN, that is, when a subject is classified in a class different than this, even if in the incorrect class, clinically it is very probable that the subject is not in a healthy state. For the AD $^{(+)}$, the system has higher Sens values than Spec, which means that when a subject is classified in a group different than AD it is very likely that such a subject does not have AD. Mean Sens values of 93% were attained and 85% for Spec. For MCI $^{(+)}$, the system has a performance similar to the AD $^{(+)}$, with 93% value for Sens and 85% for Spec.

The system has a similar behaviour with the 12-month interval data, for that reason the previous comments also apply to this setting.

These plots confirm what has been observed for the binary classification tasks, i.e, that accuracy increases with increasing numbers of time points and with the length of time intervals. They also show, that in general, median and average intensity obtain similar results whereas variance performs worse than median or average. However, if in a clinical setting the Acc is not the most important metric, but Sens or Spec are, then the variance-based features could be chosen, given that in some settings their Sens or Spec values are significantly better than the ones with average and median-based features.

The system is a probabilistic model with equal prior probability values for all the classes, these values can be optimized in order to achieve greater performance metrics of clinical interest. In this work, that optimization was not carried out because the goal was to study the system's diagnostic performance, and that work in such a small and controlled dataset would obviously create a bias in the results.

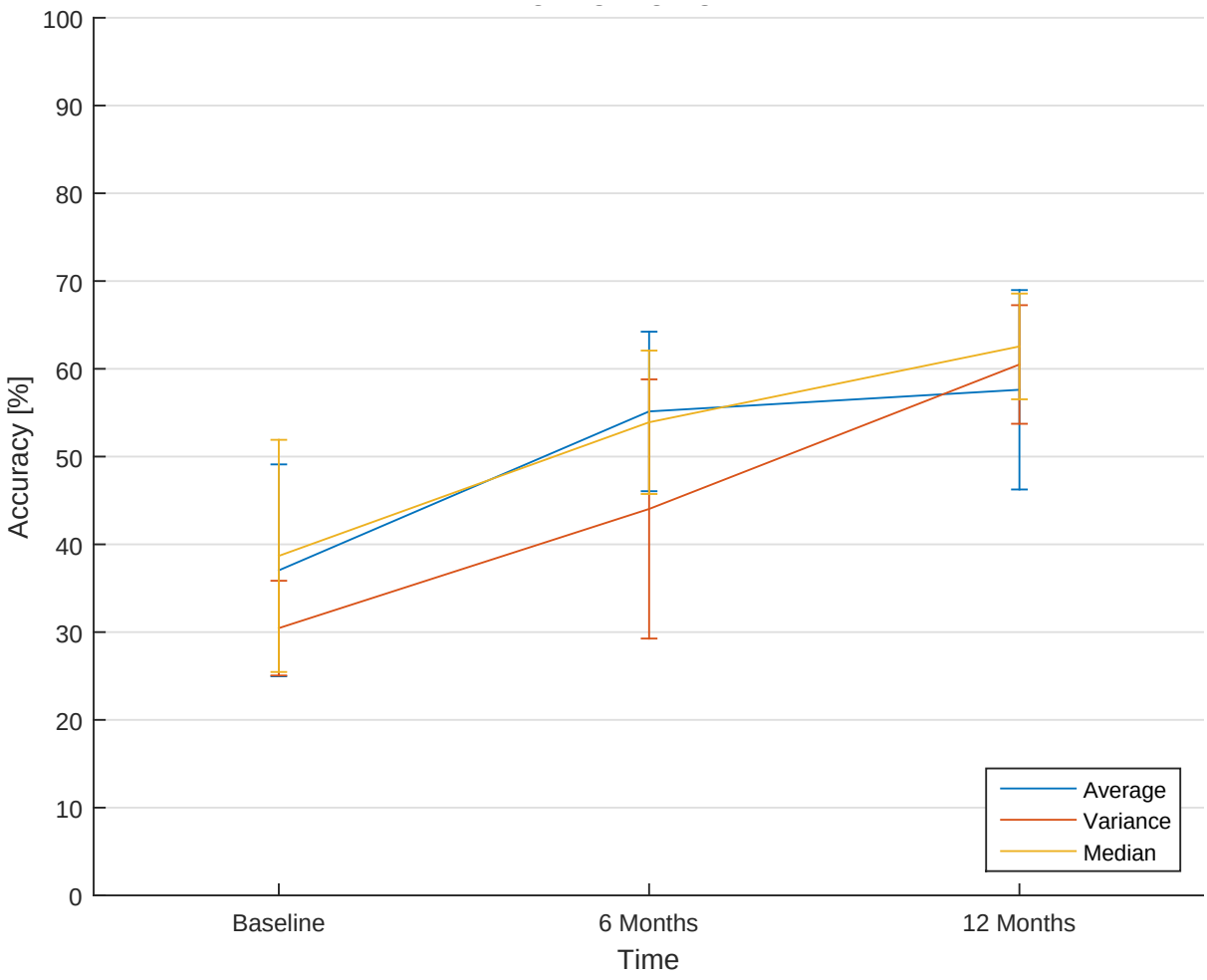
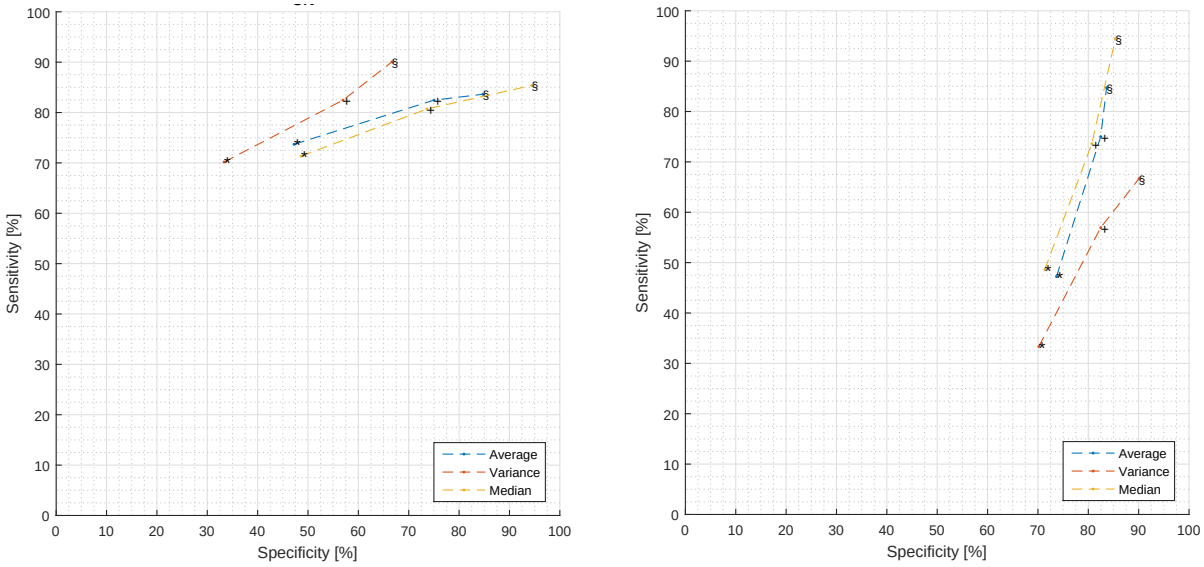
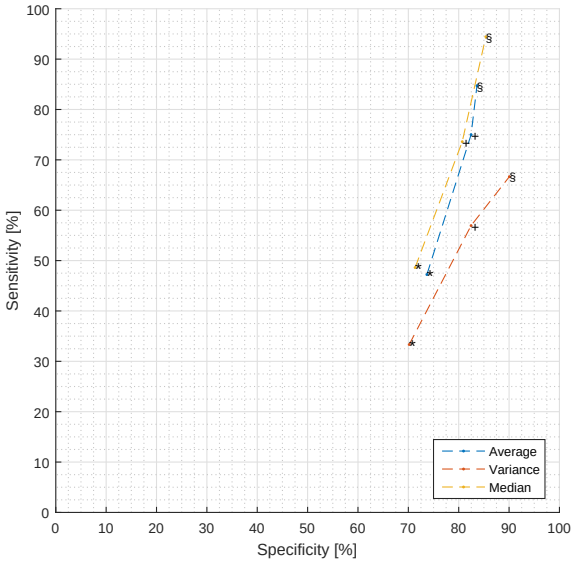


Figure 5.7: Accuracy obtained with different features for the multiclass classification task, considering different time points with a 6-month interval between follow-up scans.



(a) CN $^{(-)}$

(b) AD $^{(+)}$



(c) MCI $^{(+)}$

Figure 5.8: Multiclass Sensitivity versus Specificity, in each of the possible class couplings, for the different feature extraction schemes with a 6-month follow-up interval. The specified class in the subtitle indicates which class was considered separate from the others and whether it was assigned as positive $^{(+)}$ or negative $^{(-)}$. The result at baseline exam is represented by *, after 6 months by +, and after 12 months by §.

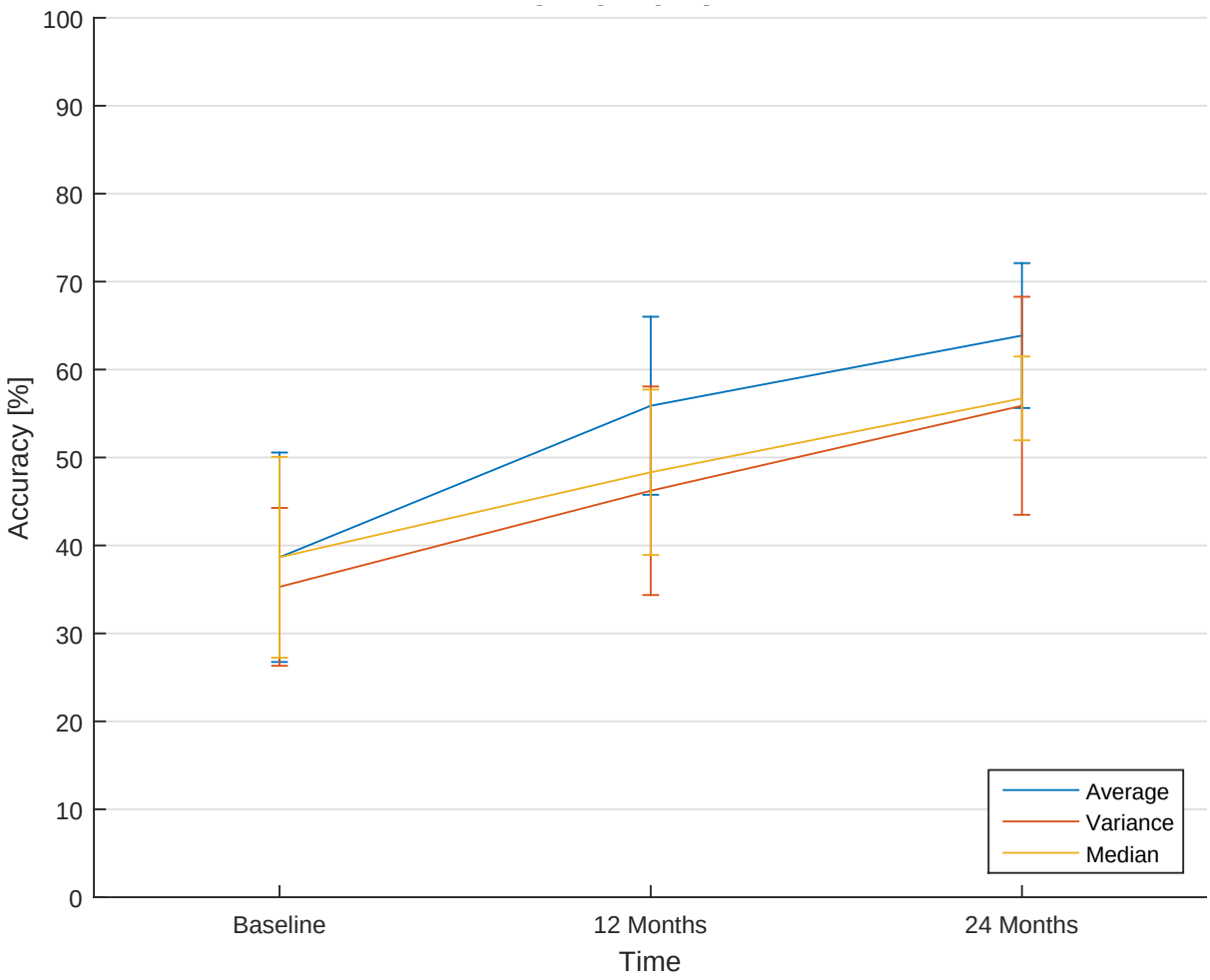
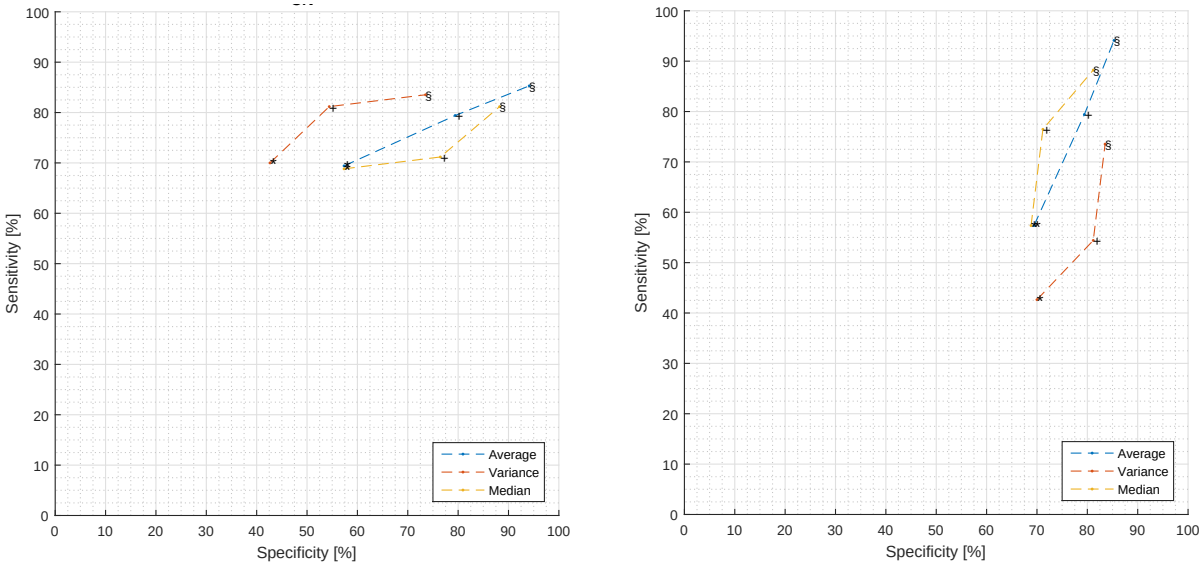
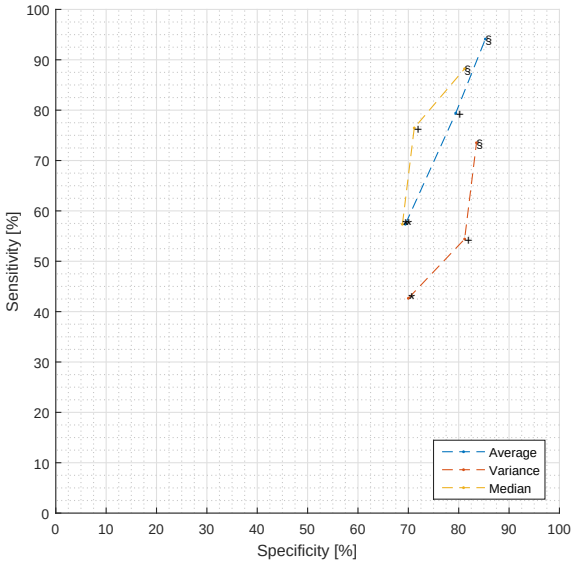


Figure 5.9: Accuracy obtained with different features for the multiclass classification task, considering different time points with a 12-month interval between follow-up scans.



(a) CN $^{(-)}$

(b) AD $^{(+)}$



(c) MCI $^{(+)}$

Figure 5.10: Multiclass Sensitivity versus Specificity, in each of the possible class couplings, for the different feature extraction schemes with a 12-month follow-up interval. The specified class in the subtitle indicates which class was considered separate from the others and whether it was assigned as positive $^{(+)}$ or negative $^{(-)}$. The result at baseline exam is represented by *, after 12 months by +, and after 24 months by §.

Chapter 6

Conclusions and Future Work

The primary goal of this work was the development of a CAD system for the diagnosis of Alzheimer's Disease at different stages which explored temporal evolution. For this purpose we proposed models for temporal evolution based on the Hidden Markov Model framework which were trained in a supervised manner using PET data from AD, MCI and CN subjects with scans taken at different time points. Our goal was accomplished since the proposed models were able to discriminate between the different stages of the disease and outperformed classification models which did not take temporal evolution into account. Besides, this model can be used from the first clinical appointment allowing an arbitrary number of updates, which is a clear advantage over systems that use a predefined number of instants.

In general, with the median and average-based features the system performed best, with a significant performance decrease when using the variance-based features. However, there are settings where this feature, although having worse Acc values, compensates with better Sens or Spec, which can be more important on a specific clinical set. Taking this in consideration, it would be interesting to classify the patients taking into account the probabilities given by model and having some sort of decision by mixing the scores form each model trained with the different features.

In order to optimize a clinical metric of interest, the classes' prior values, which were assumed equal, can be also included in the learning stage.

The fact that the MCI class is the most difficult to identify among the ones considered, leading to a poorer classification performance, indicates that this class is too heterogeneous and that further subdivions (i.e. considering more granular disease stages) of this class would benefit the system's training and performance.

The proposed models should, however, be further validated with more datasets and different image modalities. These models may also be evaluated with different image features and different techniques of dimensionality reduction, including discriminative techniques.

Overall, this method's principles are viable for clinical usage because they pose a clear view of the underlying classification processes based on a physiological reasoning, contrary to black-box or very abstract models.

In this work, the HMM models were used for the diagnosis of a subject's cognitive state. In future

developments, this approach could be used to have a unified disease model under a single HMM, adding prognostic information related to disease's progression to the clinical diagnosis.

Finally, since HMM's have the limitation that the time series have to evenly spaced, we propose to extend this work by evaluating techniques that deal with unevenly spaced time-series.

References

H. Aidos, A. Fred, A. D. N. Initiative, et al. Discrimination of alzheimer's disease using longitudinal information. *Data Mining and Knowledge Discovery*, pages 1–25, 2017.

Alzheimer's Association. 2017 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 13(4): 325–373, 2017.

J. A. Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

K. Chen, J. B. Langbaum, A. S. Fleisher, N. Ayutyanont, C. Reschke, W. Lee, X. Liu, D. Bandy, G. E. Alexander, P. M. Thompson, et al. Twelve-month metabolic declines in probable alzheimer's disease and amnestic mild cognitive impairment assessed using an empirically pre-defined statistical region-of-interest: findings from the alzheimer's disease neuroimaging initiative. *Neuroimage*, 51(2):654–664, 2010.

Y. Chen and T. D. Pham. Development of a brain mri-based hidden markov model for dementia recognition. *Biomedical engineering online*, 12(1):S2, 2013.

M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff, R. G. Thomas, R. Raman, A. C. Gamst, L. A. Beckett, C. R. Jack, M. W. Weiner, J.-F. Dartigues, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer's & Dementia*, 10(5):S400–S410, 2014.

J. Dukart, K. Mueller, A. Horstmann, B. Vogt, S. Frisch, H. Barthel, G. Becker, H. E. Möller, A. Villringer, O. Sabri, et al. Differential effects of global and cerebellar normalization on detection and differentiation of dementia in fdg-pet studies. *Neuroimage*, 49(2):1490–1495, 2010.

M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

K. R. Gray, R. Wolz, R. A. Heckemann, P. Aljabar, A. Hammers, D. Rueckert, A. D. N. Initiative, et al. Multi-region analysis of longitudinal fdg-pet for the classification of alzheimer's disease. *Neuroimage*, 60(1):221–229, 2012.

M. Huang, W. Yang, Q. Feng, W. Chen, A. D. N. Initiative, et al. Longitudinal measurement and hierarchical classification framework for the prediction of alzheimer's disease. *Scientific reports*, 7, 2017.

D. Izquierdo-Garcia, A. E. Hansen, S. Förster, D. Benoit, S. Schachoff, S. Fürst, K. T. Chen, D. B. Chonde, and C. Catana. An spm8-based approach for attenuation correction combining segmentation and non-rigid template formation: application to simultaneous pet/mr brain imaging. *Journal of nuclear medicine: official publication, Society of Nuclear Medicine*, 55(11):1825, 2014.

I. T. Jolliffe. Principal component analysis and factor analysis. pages 115–128, 1986.

Y.-E. S. Ju, J. S. McLeland, C. D. Toedebusch, C. Xiong, A. M. Fagan, S. P. Duntley, J. C. Morris, and D. M. Holtzman. Sleep quality and preclinical alzheimer disease. *JAMA neurology*, 70(5):587–593, 2013.

A. Küntzelmann, T. Guenther, U. Haberkorn, M. Essig, F. Giesel, R. Henze, M. L. Schroeter, J. Schröder, and P. Schönknecht. Impaired cerebral glucose metabolism in prodromal alzheimer's disease differs by regional intensity normalization. *Neuroscience letters*, 534:12–17, 2013.

T. Y. Lam and I. M. Meyer. Efficient algorithms for training the parameters of hidden markov models using stochastic expectation maximization (em) training and viterbi training. *Algorithms for Molecular Biology*, 5(1):38, 2010.

J. A. Maldjian, P. J. Laurienti, R. A. Kraft, and J. H. Burdette. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fmri data sets. *Neuroimage*, 19(3):1233–1239, 2003.

N. Mattsson, H. Zetterberg, O. Hansson, N. Andreasen, L. Parnetti, M. Jonsson, S.-K. Herukka, W. M. van der Flier, M. A. Blankenstein, M. Ewers, et al. Csf biomarkers and incipient alzheimer disease in patients with mild cognitive impairment. *Jama*, 302(4):385–393, 2009.

E. McDade and R. J. Bateman. Stop alzheimer's before it starts. *Nature*, 547(7662):153, 2017.

M. Mintun, G. Larossa, Y. Sheline, C. Dence, S. Y. Lee, R. Mach, W. Klunk, C. Mathis, S. DeKosky, and J. Morris. [11c] pib in a nondemented population potential antecedent marker of alzheimer disease. *Neurology*, 67(3):446–452, 2006.

P. M. Morgado, M. Silveira, A. D. N. Initiative, et al. Minimal neighborhood redundancy maximal relevance: Application to the diagnosis of alzheimer's disease. *Neurocomputing*, 155:295–308, 2015.

T. D. Pham, F. Salvetti, B. Wang, M. Diani, W. Heindel, S. Knecht, H. Wersching, B. T. Baune, and K. Berger. The hidden-markov brain: comparison and inference of white matter hyperintensities on magnetic resonance imaging (mri). *Journal of neural engineering*, 8(1):016004, 2011.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

M. N. Samtani, M. Farnum, V. Lobanov, E. Yang, N. Raghavan, A. DiBernardo, and V. Narayan. An improved model for disease progression in patients from the alzheimer's disease neuroimaging initiative. *The Journal of Clinical Pharmacology*, 52(5):629–644, 2012.

P. Scheltens, K. Blennow, M. M. B. Breteler, B. de Strooper, G. B. Frisoni, S. Salloway, and W. M. V. der Flier. Alzheimer's disease. *The Lancet*, 388(10043):505 – 517, 2016. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(15)01124-1. URL `http://www.sciencedirect.com/science/article/pii/S0140673615011241`.

R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using hidden markov models. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2845–2848. IEEE, 2012.

P. J. Visser, F. Verhey, D. L. Knol, P. Scheltens, L.-O. Wahlund, Y. Freund-Levi, M. Tsolaki, L. Minthon, Å. K. Wallin, H. Hampel, et al. Prevalence and prognostic value of csf markers of alzheimer's disease pathology in patients with subjective cognitive impairment or mild cognitive impairment in the descripa study: a prospective cohort study. *The Lancet Neurology*, 8(7):619–627, 2009.

G. Waldemar, B. Dubois, M. Emre, J. Georges, I. McKeith, M. Rossor, P. Scheltens, P. Tariska, and B. Winblad. Recommendations for the diagnosis and management of alzheimer's disease and other disorders associated with dementia: Efns guideline. *European Journal of Neurology*, 14(1), 2007.

B. Wang and T. D. Pham. Mri-based age prediction using hidden markov models. *Journal of neuroscience methods*, 199(1):140–145, 2011.

W.-H. Wang, H.-L. Wu, P.-C. Chung, and M.-C. Pai. An hmm-based gait comparison: Using alzheimer's disease patients as examples. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–6. IEEE, 2015.

Y. Wang, S. M. Resnick, and C. Davatzikos. Spatio-temporal analysis of brain mri images using hidden markov models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 160–168. Springer, 2010.

I. Yakushev, A. Hammers, A. Fellgiebel, I. Schmidtmann, A. Scheurich, H.-G. Buchholz, J. Peters, P. Bartenstein, K. Lieb, and M. Schreckenberger. Spm-based count normalization provides excellent discrimination of mild alzheimer's disease and amnestic mild cognitive impairment from healthy aging. *Neuroimage*, 44(1):43–50, 2009.