Alzheimer's Disease Modelling through Hidden Markov Models

Diogo Mendes Cardoso
diogo.m.cardoso@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa

Abstract

In the last few decades there has been a great amount of work in the development of Computer-Aided Diagnosis (CAD) that can diagnose Alzheimer's disease as early and as accurately as possible, from neuroimaging data such as Magnetic Resonance Imaging (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 Model (HMM) for modelling the evolution of Alzheimer's Disease (AD), Mild Cognitive Impairement (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, PET images, Hidden Markov Models, Disease Modelling.

1 Introduction

1.1 Motivation

According to [1], 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 MRI and PET. Together with 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 longitudinal data or the time evolution. 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 State of the Art

Although many computer aided diagnosis systems (CAD) have been proposed for the diagnosis of Alzheimer's disease from neuroimaging data, 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 [2], [3] or [4] which 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 [5], [6] and [7].

1.2.1 Modelling temporal evolution

In [8], a method based on HMM was proposed to model AD progression. 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 on four feature vectors that were extracted from MRI Alzheimer's Disease Neuroimaging Initiative (ADNI) 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). It is assumed that each hidden state accounts for a correspondent disease counterpart which implies a left-to-right hidden state transition structure. With the small nuance that each state is allowed to go back one state, accounting for possible disease reversions. 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 clusters in this left-to-right model corresponded effectively to a disease progression one, Viterbi decoding was done for each sequence in order to see to which state that 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 classes, 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.

In [9] another promising work using HMM with MRI scans was presented, but unlike 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 vectors 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%, Sensibility (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% Acc when comparing AD with middle-aged and 98.7% with young.

1.3 Proposed Approach

We propose to build a CAD system for AD at different stages that allows the use of follow-up information to complement its decision when needed. For this, we propose to develop separate HMM for modelling the evolution of AD, MCI and CN individuals.

1.4 Original Contributions

Although widely used in other areas such as speech processing, statistical modelling of image data using HMM was barely used, only a couple of times with follow-up information and, to our knowledge, never with PET. This methodology is promising and allows for simultaneous diagnosis and prognosis.

1.5 Outline

The present thesis is organized in the following way:

- In Section 2, we describe the CAD system proposed to diagnose and model the temporal evolution of AD with PET.
- In Section 3.4, we present a description of the data used in the experiments and the systems' performance is evaluated and discussed.
- Finally, in Chapter 4, we present the thesis conclusions and suggestions for future work.

2 Methods

2.1 Proposed Approach

2.1.1 Feature Extraction

In the context of AD, there are specific brain regions where the disease has a greater impact [10]. 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.

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

After feature extraction, Principal Components Analysis (PCA) [11] 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.

2.1.2 Model Learning

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. 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 also probabilistically defined by a distribution that describes the possible transitions. [12].

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 1 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 infeasible to train a 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 feature extraction scheme presented in the previews section.

While describing a 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.

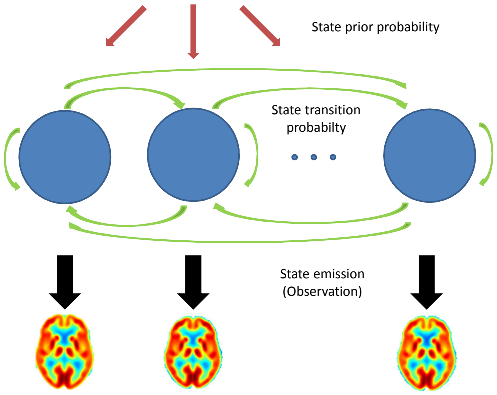


Figure 1: HMM-based disease model.

In this problem, the emissions where considered continuous following a Gaussian distribution in each hidden state, with full-covariance matrices.

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 relates directly with 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 [13].

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.

Having defined the number of hidden states and all parameters starting points, the model's parameters are estimated using the Expectation-Maximization (EM) algorithm for HMM [14].

2.1.3 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 1).

$$P(x|C) \tag{1}$$

Considering $f(x)$ to be the intended classifier's output for the observed sequence x, represented in equation 2. However under the model we cannot determine directly such quantities, using Bayes' rule, equation 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.

$$f(x) \in \arg\max_{C} P(C|x) \tag{2}$$

$$f(x) \in \arg\max_{C} \frac{P(x|C)P(C)}{p(x)} \tag{3}$$

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

$$f(x) \in \arg\max_{C} P(x|C) \tag{4}$$

The scheme represented in Fig. 2 summarizes the classification process, given a PET scan sequence of arbitrary temporal length.

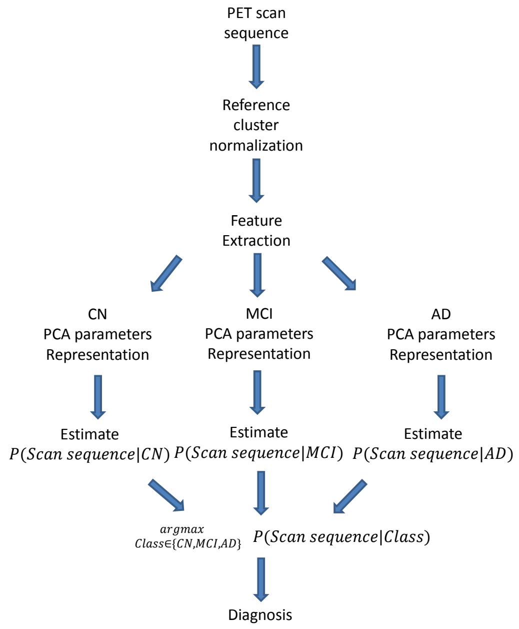

PET scan sequence
Reference cluster normalization
Feature Extraction
CN PCA parameters Representation
MCI PCA parameters Representation
AD PCA parameters Representation
Estimate $P(Scan\ sequence|CN)$
Estimate $P(Scan\ sequence|MCI)$
Estimate $P(Scan\ sequence|AD)$
$\underset{Class \in \{CN,MCI,AD\}}{argmax}$ $P(Scan\ sequence|Class)$
Diagnosis

Figure 2: Classification protocol.

3 Results and Discussion

3.1 Dataset

The data used in this work was retrieved from 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 is constituted by 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 at 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 1.

3.2 Preprocessing

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

3.2.1 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

Table 1: Clinical (Mini Mental State Examination (MMSE) Clinical Dementia Rate (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

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 [15], 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 [16]. Tissue classification, on the other hand, was conducted with SPM8 [17] 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 [17]. 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. [18] contains a more detailed description of the registration protocol here described.

3.2.2 Intensity Normalization

Intensity normalization is also necessary since it greatly influences the classification process [19, 20].

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 [21]. 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 [21], 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.

3.3 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 consists on dividing the data into 10 folds with similar size. Then, in each iteration, one of the folds is selected to test (testing fold) and 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 multi-class performance should be computed. However, it is also common in CAD systems to reduce the multi-class 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).

3.4 Classification Results

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

3.5 Binary Classification

As seen in chapter 2, 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.

Analysing the classification accuracy (mean and variance) obtained at baseline for varying numbers of principal components used, the overall accuracy is greater in the CN vs. AD task, being 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 it demonstrates to be in line with the other feature extraction schemes performance. At baseline, the average and the median in the defined Region of Interest (ROI) demonstrate to be valid features to the addressed problem of automated classification of patients in the possible AD stages. However, any 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, once again, the best overall results concern the CNvs. 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 CNvs. MCI and MCIvs. AD.

Figure 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 but for 12 month intervals between follow-up scans. The results in both figures were obtained with 10 PCA components.

The plots in Figure 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 to 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 4, it is unequivocal that the inclusion of time information improves the systems Sens and Spec, reaching to 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 show that, as expected, accuracy increases with increasing number of time points and that the variability across folds diminishes (smaller error bars), as in the 6 month-interval case. The CN vs. MCI task is the one with the lowest Acc values, but in this case, with 12 month-intervals, the variance-based features approach the performance of the average and media-based features. In two years interval, with two exams, the system reaches to an Acc value of 99% for CN vs. MCI, 75% for CN vs. MCI, and 89% for MCI vs. AD.

Analysing the Sens and Spec plots in figure 4, it unequivocal that the inclusion of time information improves the systems Sens and Spec, as in the previous case, reaching to 100% Sens and 95% Spec for CN vs. AD, 85% Sens and 85% 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.

From the comparison of figures 3 and 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.bMCI 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 , leading to slightly different datasets, which could make a difference in such a broad and heterogeneous class as MCI.

3.6 Multiclass Classification

Figures 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 and 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 a 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 these cases the ones that represent the higher 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 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 is very probable that the subject is not in a healthy state. For the AD $^{(+)}$, the system has higher Sens values than Spec, meaning that when a subject is classified in a group different than AD it is very likely that such subject does not have AD. It reaches mean Sens values of 93% and 85% for Spec. For MCI $^{(+)}$, the system has a performance similar to the AD $^{(+)}$, with 93% value for Sens and 85% for Spec.

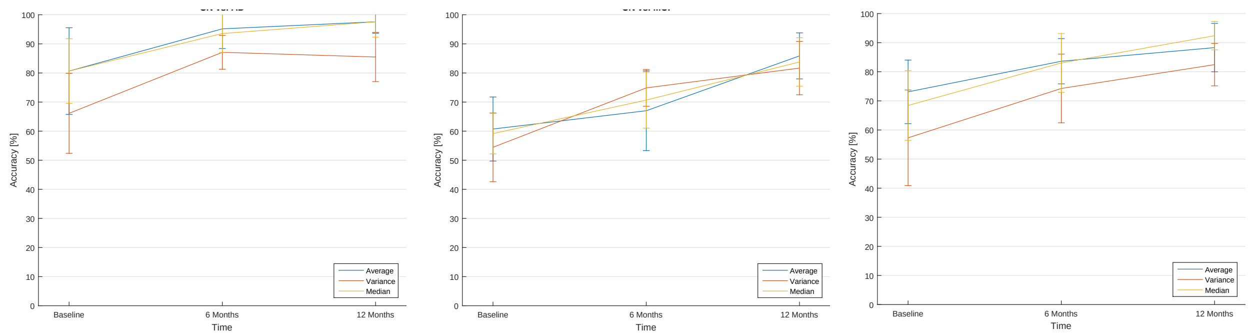
The system has a similar behaviour with the 12-month interval data, for that reason the previous comments also apply to this setting.

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 that 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 one 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 it was intended to study the system's class-separability performance, and that work in such a small and controlled dataset would obviously create a bias in the results.

4 Conclusion

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 tem-

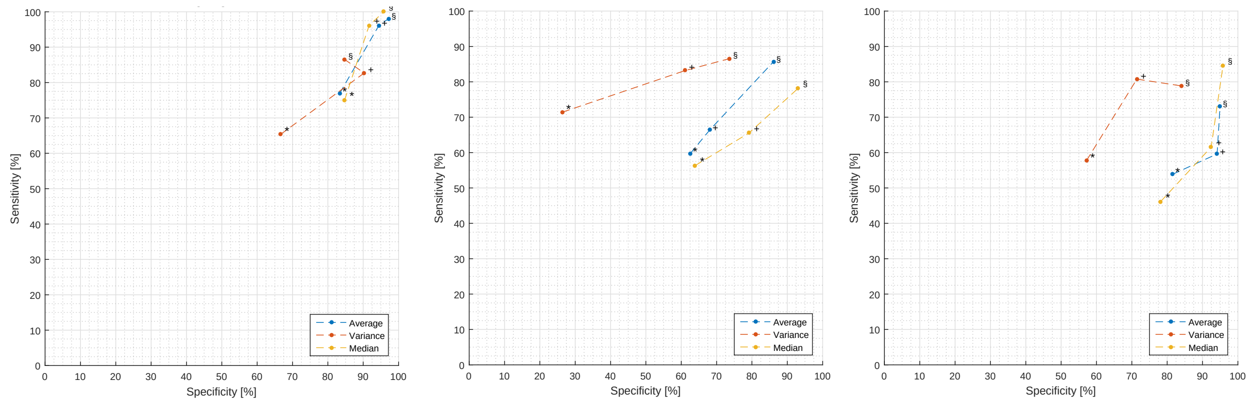


(a) CN vs. AD

(b) CN vs. MCI

(c) MCI vs. AD

Figure 3: Accuracy obtained with different features for each binary classification task, considering different time points with a 6-month interval between follow-up scans.

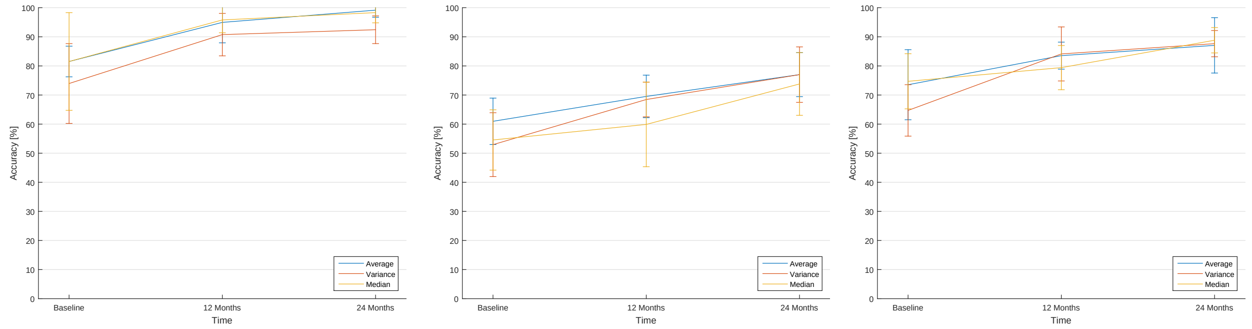


(a) CN vs. AD

(b) CN vs. MCI

(c) MCI vs. AD

Figure 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 §.

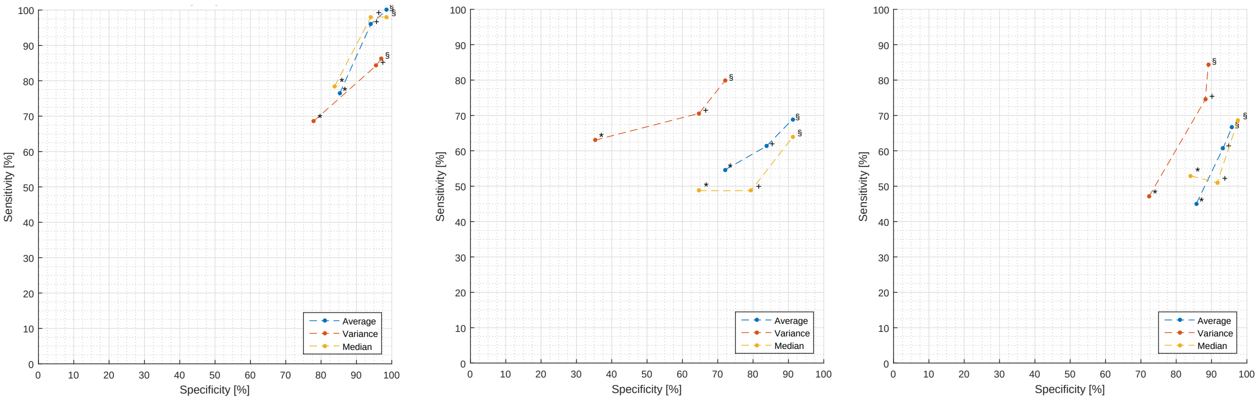


(a) CN vs. AD

(b) CN vs. MCI

(c) MCI vs. AD

Figure 5: Accuracy obtained with different features for each binary classification task, considering different time points with a 12-month interval between follow-up scans.

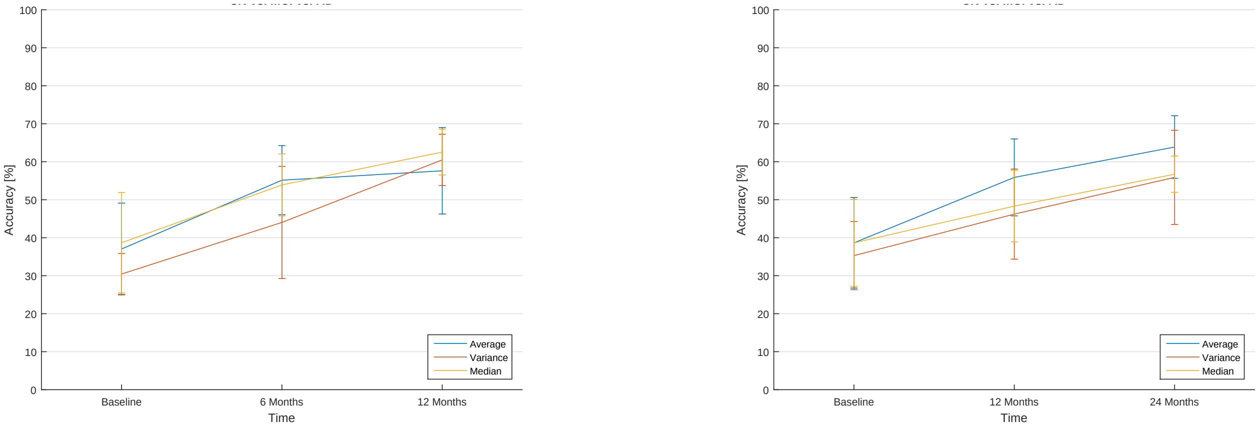


(a) CN vs. AD

(b) CN vs. MCI

(c) MCI vs. AD

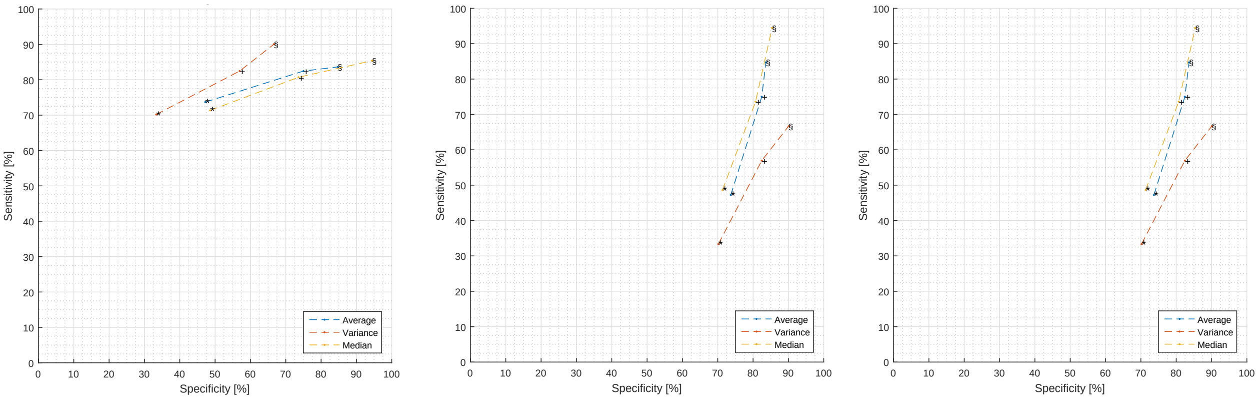
Figure 6: Sensitivity versus Specificity obtained with different features for each binary classification task and considering a 12-month follow-up interval. The result at baseline exam is represented by *, after 12 months by +, and after 24 months by §.



(a) 6 month interval

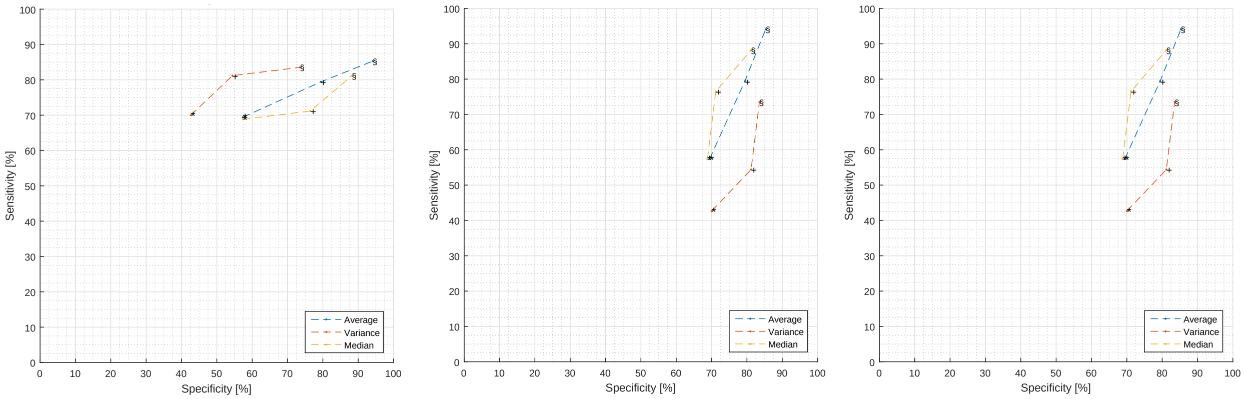
(b) 12 month interval

Figure 7: Accuracy obtained with different features for the multiclass classification task, considering different time points and 6 and 12 month intervals



(a) CN $^{(-)}$ (b) AD $^{(+)}$ (c) MCI $^{(+)}$

Figure 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 §.



(a) CN $^{(-)}$ (b) AD $^{(+)}$ (c) MCI $^{(+)}$

Figure 9: 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 §.

poral 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.

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.

In this work, the HMM models were used for the diagnosis of a subject's cognitive state. However, they can also provide prognostic information related to disease's progression. 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

[1] Alzheimer's Association, "2017 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 13, no. 4, pp. 325–373, 2017.

[2] 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*, vol. 60, no. 1, pp. 221–229, 2012.

[3] 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*, vol. 7, 2017.

[4] H. Aidos, A. Fred, A. D. N. Initiative *et al.*, "Discrimination of Alzheimer's disease using longitudinal information," *Data Mining and Knowledge Discovery*, pp. 1–25, 2017.

[5] 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*, vol. 51, no. 2, pp. 654–664, 2010.

[6] 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*, vol. 10, no. 5, pp. S400–S410, 2014.

[7] 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*, vol. 52, no. 5, pp. 629–644, 2012.

[8] 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.* IEEE, 2012, pp. 2845–2848.

[9] Y. Chen and T. D. Pham, "Development of a brain mri-based Hidden Markov model for dementia recognition," *Biomedical engineering online*, vol. 12, no. 1, p. S2, 2013.

[10] J. T. Coyle, D. L. Price, and M. R. Delong, "Alzheimer's disease: a disorder of cortical cholinergic innervation," *Science*, vol. 219, no. 4589, pp. 1184–1190, 1983.

[11] I. T. Jolliffe, "Principal component analysis and factor analysis," pp. 115–128, 1986.

[12] L. R. Rabiner, "A tutorial on Hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[13] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[14] 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*, vol. 4, no. 510, p. 126, 1998.

[15] 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*, vol. 19, no. 3, pp. 1233–1239, 2003.

[16] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[17] 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*, vol. 55, no. 11, p. 1825, 2014.

[18] 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*, vol. 155, pp. 295–308, 2015.

[19] 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*, vol. 49, no. 2, pp. 1490–1495, 2010.

[20] 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*, vol. 534, pp. 12–17, 2013.

[21] 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*, vol. 44, no. 1, pp. 43–50, 2009.